Elements of Information Theory

Bilingual course
(Chinese taught course)
Information and Communication Eng. Dept.
Deng Ke

Outline

- Entropy
- Source
- Joint entropy and conditional entropy
- Chain rule
- Channel
- Mutual information
- Some Inequalities
- Summary

Entropy

- Entropy is associated with a set, such as source
- Entropy is the first-order statistics of the information, is the mathematical expectation
- Example, the simplest source, can only send 0 and 1
  - Source 1: $P_0 = P_1 = 0.5$
  - Source 2: $P_0 = 0.1 \quad P_1 = 0.9$

Entropy

- ***Definition:*** The *entropy* $H(X)$ of a discrete random variable X is defined by

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

- X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function

$$p(x) = \Pr(X = x), x \in \mathcal{X}$$

- Expectation interpretation

$$H(X) = E_p\left[\log\frac{1}{p(X)}\right] = -\sum p(x)\log p(x)$$

- Entropy is the expectation of the information

Entropy

- The understanding of entropy
  - H_1=1 bit, H_2=0.469 bit $H_1 > H_2$
  - entropy is associated with the number of the microstates of a system. Obviosly, source 1 has more microstates
- For a given n, the maximum $H(X)$ is equal to $\log n$, when $p(x_i)=1/n, \forall i$
- First, we demonstrate: $\ln z \le z - 1 \quad (z > 0)$

"=" if and only if $z = 1$

Consider $f(z) = \ln z - (z-1)$, All we need are $f(z) \leqslant 0$

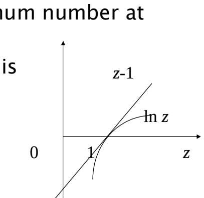
Entropy

- And we have

$$f'(z) = 1/z - 1 \qquad f''(z) = -1/z^2 < 0$$

- Then $f(z)$ has the maximum number at $f'(z)=0$, i.e. $z=1$
- The maximum number is

$$f(z) \le f(z)|_{z=1} = \ln 1 - (1-1) = 0$$



Entropy

$$H(X)-\log K=\sum_i p(x_i)\log\frac{1}{p(x_i)}-\sum_i p(x_i)\log K$$
$$=\log e\sum_i p(x_i)\ln\frac{1}{p(x_i)K}$$
$$\le \log e\sum_i p(x_i)\left(\frac{1}{p(x_i)K}-1\right)$$
$$=\log e\left[\sum_i\frac{1}{K}-\sum_i p(x_i)\right]=\log e(1-1)=0$$

Entropy

- $K\geq 2$
- When K=2, let $P(1)=p$, then $P(0)=1-p$

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

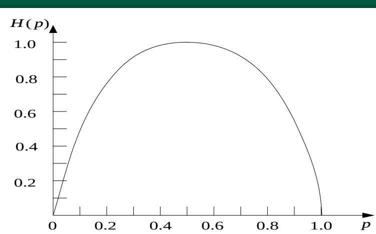
- Denote $H(X)=\Omega(p)$
- Actually, the notion of entropy comes from Thermo-dynamics.

$$S=k\,(\ln\Omega)$$

Boltzmann's constant

Number of random microstates

Entropy

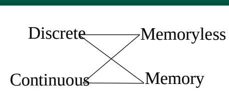


Entropy

- Unit conversion

  1nat=$\log_2 e$ bit $\approx$1.44bit   1bit=ln2 nat $\approx$0.69nat
- The entropy of a fair coin toss is 1 bit
  - The origin of "bit" in computer science
- Example
- Let $X=\begin{cases} a & p(a)=1/2 \\ b & p(b)=1/4 \\ c & p(c)=1/8 \\ d & p(d)=1/8 \end{cases}$
- Then $H(X)=-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{8}\log\frac{1}{8}-\frac{1}{8}\log\frac{1}{8}=1.75\text{bits}$

Source



- The Kind of Source
- Explanation
  - Discrete: Finite possibilities, finite number of elements
  - Memoryless: The statistics independence of sending symbols
- Example

Source

- 10 black balls and 10 white balls in a bag. once a ball, put it back, memoryless; do not put it back, memory
- The entropy of a discrete memoryless source

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

- The describe of discrete memory source requires the conditional entropy and the joint entropy

Source Coding

- Text
  - ASCII , 128 symbols, 7 bits
  - GB2312, 6763 characters, at least 13 bits, actually 14 bits, 2 bytes (2^{13}=8192>6763> 2^{12}=4096)
  - GBK, GB2312+BIG5, >30000 characters, occupies 15 bits (2^{15}=32768)
  - Unicode
  - UTF-8
  - UTF-16

Source Coding

- Voice
  - CD
  - MP3
- Image
  - JPG
- Video
  - MPEG1 VCD 1.5M bps
  - MPEG2 DVD 12M bps
  - RMVB 225K, 350K, 450K bps

Joint Entropy

- Joint Entropy
  - Extend the definition form a single random variable to a pair of random variables
  - Single is simple; double is trouble; triple is terrible.
- ***Definition:*** The *joint entropy* $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

Joint Entropy

- which can also be expressed as

$$H(X,Y) = -E[\log p(x,y)]$$

- In this definition (X,Y) can be considered to be a single vector-valued random variable.
- Conditional entropy
  - Define the conditional entropy of a random variable given another as --the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

Conditional Entropy

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

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &= -\sum_x \sum_y p(x,y) \log p(y \mid x) \\ &= -E \log p(Y \mid X) \end{aligned}$$

Example

- Let one discrete memory source has the probability

$$\begin{matrix} \mathbf{X}: & a_1, & a_2, & a_3 \\ P(\mathbf{X}): & \frac{11}{36} & \frac{4}{9} & \frac{1}{4} \end{matrix}$$

- and the conditional probability $P(a_j|a_i)$,

	a_j		
a_i	a_1	a_2	a_3
a_1	9/11	2/11	0
a_2	1/8	3/4	1/8
a_3	0	2/9	7/9

Please compute $H(X^2)$

Example

- According to the definition of joint entropy

$$H(\mathbf{X}^2) = -\sum_{i=1}^{3}\sum_{j=1}^{3} P(a_i a_j)\log P(a_i a_j)$$

- The joint probability can be computed as

$P(a_1a_1)= P(a_1)\, P(a_1|a_1)=(11/36)\times(9/11)=1/4$

$P(a_1a_2)= P(a_1)\, P(a_2|a_1)=(11/36)\times(2/11)=1/18$

⋮

$P(a_3a_3)= P(a_3)\, P(a_3|a_3)=(1/4)\times(7/9)=7/36$

$H(\mathrm{X}^2)$=2.412 bits

- Also, we have

Example

$$H(\mathbf{X}) = -\sum_{i=1}^{3} P(a_i)\log P(a_i) = 1.542 \quad \text{bits/symbo l}$$

$$H(X_2 \mid X_1) = -\sum_{i=1}^{3}\sum_{j=1}^{3} P(a_i a_j)\log P(a_j \mid a_i) = 0.870 \quad \text{bits/symbo l}$$

- **$H(\mathrm{X})+ H(X_2|X_1)$ = 2.412 bits**
- **Now, $H(\mathrm{X}^2) = H(\mathrm{X})+H(X_2|X_1)$**
- **$H(\mathrm{X}) > H(X_2|X_1)$**
- **$H(\mathrm{X}^2) < 2H(\mathrm{X})$**

Discrete Memory Source

- K-order memory: If the symbol correlates with K transmitted symbols, the source is K-order discrete memory source.
- 1-order memory: $P(U|Q) \approx 1$
- Two conclusions from the example
  - Chain rule
    - **$H(\mathrm{X}^2) = H(\mathrm{X})+H(X_2|X_1)$**
  - Memory means redundancy, leads to the loss of information
    - **$H(\mathrm{X}^2) < 2H(\mathrm{X})$**

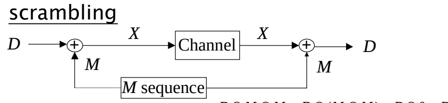
Chain rule

- The entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other.
- Joint entropy = entropy + conditional entropy
- joint probability = marginal probability * conditional probability $p(x, y)=p(x)p(y|x)$
- Theorem 2.2.1 (Chain rule)

$$H(X,Y) = H(X) + H(Y \mid X)$$

Scramble

- In real communications, memory source can be transformed to memoryless source, with scrambling



$$D \oplus M \oplus M = D \oplus (M \oplus M) = D \oplus 0 = D$$

- $P(X=1)=P(D=0,M=1)+P(D=1,M=0)=P(D=0)*1/2+P(D=1)*1/2=1/2$
- $P(X=0)=1-P(X=1)=1/2$

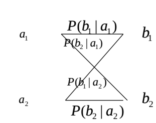
Channel Model

- The Model of Binary Discrete Memoryless Channel （ BDMC ）

The explaination of forward transition probability

$P(b_1|a_1)$ The probability of sending **a_1** and receiving **b_1**

$P(b_1|a_2)$ The probability of sending **a_2** and receiving **b_1** and so on......



Channel Model

- Example 1：Binary Symmetric Channel :BSC

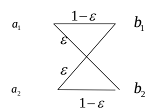
$P(b_1|a_1)=P(b_2|a_2)=1-\varepsilon$ ε:error rate

$P(b_1|a_2)=P(b_1|a_2)=\varepsilon$

When ε=1/2，the input is independent with the output，completely-noisy-channel(CNC)

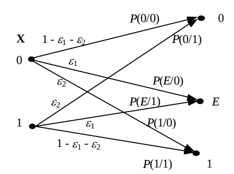
When ε=0，noiseless channel

CNC can not transmit the information

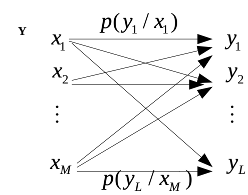


Channel Model

- Ex. 2: Binary erasure channel



Ex. 3: Common model



Mutual Information

- Meaning: The information of once transmission
- Some equations

Let the channel input $X\in\{a_1,a_2,\ldots,a_K\}$ ，the channel output $Y\in\{b_1,b_2,\ldots,b_J\}$ ，The joint probablity $P(a_k,b_j)$。Then the input $P(a_k)=\sum_j P(a_k,b_j)$

Output： $P(b_j)=\sum_k P(a_k,b_j)$

Forward transition： $P(b_j|a_k)=\dfrac{P(a_k,b_j)}{P(a_k)}$

Backward transition： $P(a_k|b_j)=\dfrac{P(a_k,b_j)}{P(b_j)}$

Mutual Information

The computation of mutual information

- If the input of channel is a_k，the information before transmission is $I(a_k)=\log\dfrac{1}{p(a_k)}$
- The output of channel is b_j，then the information after transmission about a_k is

$$I(a_k|b_j)=\log\frac{1}{p(a_k|b_j)}$$

- The information of a_k changes before and after transmission, then the transmission information is

$$I(a_k;b_j)\equiv I(a_k)-I(a_k|b_j)=\log\frac{1}{p(a_k)}-\log\frac{1}{p(a_k|b_j)}=\log\frac{p(a_k|b_j)}{p(a_k)}$$

Mutual Information

- **When the channel is noiseless**

The source sends a_k，and the sink get its all information,then

$$I(a_k;b_j)=I(a_k)$$

Mutual Information

- The average of mutual information

$$I(X;Y)\equiv \underset{XY}{E}[I(x;y)]=\sum_{k,j}p(a_k,b_j)I(a_k;b_j)=\sum_{k,j}p(a_k,b_j)\log\frac{p(a_k|b_j)}{p(a_k)}$$

- Also

$$I(X;Y)=I(Y;X)=\sum_{x,y}p(x,y)I(x;y)$$

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

- $I(x;y)$ may be positive, negative, or zero；but $I(X;Y)$ is non-negative。

Mutual Information

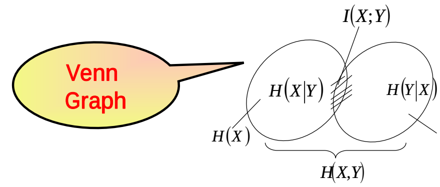
- Summary： $I(X;Y)=I(Y;X)$

$$
\begin{aligned}
&= \sum_{x,y} p(x,y) I(x;y) = \sum_{x,y} p(x,y) I(y;x) \\
&= \sum_{x,y} p(x,y)\log\frac{p(x\mid y)}{p(x)} = \sum_{x,y} p(x,y)\log\frac{p(y\mid x)}{p(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= H(X)-H(X\mid Y) \\
&= H(Y)-H(Y\mid X) \\
&= H(X)+H(Y)-H(X,Y)
\end{aligned}
$$

***H*(X，Y) is the joint entropy, and H(Y|X) is the conditional entropy.**

Mutual Information

- Two circles

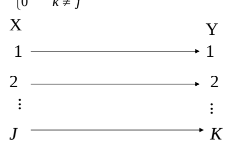


Mutual Information

- Examples of mutual information computation
- Example 1: Noiseless channel

Channel model

$$p(a_k\mid b_j)=p(b_j\mid a_k)=\begin{cases}1 & k=j\\ 0 & k\neq j\end{cases}$$



Mutual information

I(X;Y)=H(X)

Mutual Information

- Example 2: completely noisy channel

Channel model

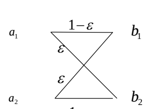
$$p(y\mid x)=p(y)$$

Mutual information I(X;Y)=0

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(y\mid x)}{p(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(y)}{p(y)} \\
&= \sum_{x,y} p(x,y)\log 1 \\
&= 0
\end{aligned}
$$

Mutual Information

- Example 3: BSC，



Please demonstrate:

$$H(Y\mid X)=\Omega(\varepsilon)$$

$$I(X;Y)=\Omega(\varepsilon+p-2\varepsilon p)-\Omega(\varepsilon)$$

where $\quad \Omega(x)=x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}$

Mutual Information

$$
\begin{aligned}
H(\mathbf{Y}\mid\mathbf{X}) &= -\sum_{\mathbf{XY}} P(xy)\log P(y\mid x) \\
&= -P(0)P(0\mid 0)\log P(0\mid 0) - P(0)P(1\mid 0)\log P(1\mid 0) \\
&\quad -P(1)P(0\mid 1)\log P(0\mid 1) - P(1)P(1\mid 1)\log P(1\mid 1) \\
&= -(1-\varepsilon)\log(1-\varepsilon)-\varepsilon\log\varepsilon = \Omega(\varepsilon)
\end{aligned}
$$

$$H(\mathbf{Y})=\Omega[p(b_1)]$$

$$
\begin{aligned}
p(b_1) &= \sum_k p(a_k,b_1) = p(a_1,b_1)+p(a_2,b_1) \\
&= p(a_1)p(b_1\mid a_1)+p(a_2)p(b_1\mid a_2) \\
&= p(1-\varepsilon)+(1-p)\varepsilon \\
&= p+\varepsilon-2\varepsilon p
\end{aligned}
$$

$$H(\mathbf{Y})=\Omega[p(b_1)]=\Omega(p+\varepsilon-2\varepsilon p)$$

Mutual Information

$\therefore\ I(X;Y)=\Omega(\varepsilon+p-2\varepsilon p)-\Omega(\varepsilon)$

- Special cases

  $\varepsilon=0$ Noiseless Channel $I(X;Y)=\Omega(0+p-0)-\Omega(0)=H(X)$

  Just like example 1

  $\varepsilon=1/2$ CNC $I(X;Y)=\Omega(1/2+p-2\bullet 1/2\bullet p)-\Omega(1/2)=\Omega(1/2)-\Omega(1/2)=0$

  Just like example 2

Xi'an Jiaotong University

Mutual Information

- The Nonnegativity of mutual information

  For any two discrete random variables X, Y,

$$I(X;Y)\geq 0$$

with equality if and only if X and Y are independent

$I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)\geqslant 0$

$H(X)\geqslant H(X|Y)$ $H(Y)\geqslant H(Y|X)$

Xi'an Jiaotong University

Some Inequalities

- Convexity
- Convex function (cup)

  Definition A function $f(x)$ is said to be *convex* over an interval (a,b) if for every $x_1, x_2\in(a,b)$ and $0\leq\lambda\leq 1$

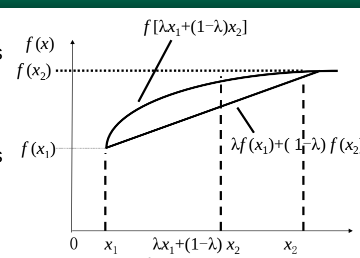
  $$f(\lambda x_1+(1-\lambda)x_2)\leq\lambda f(x_1)+(1-\lambda)f(x_2)$$

  A function f is said to be *strictly convex* if equality holds only if $\lambda=0$ or $\lambda=1$.
- Concave function (cap)
  - A function f is *concave* if $-f$ is convex.

Xi'an Jiaotong University

Concave Function

- Cup: if it always lies below any chord
- Cap: if it always lies above any chord.



cap function

Xi'an Jiaotong University

Convexity

- Examples
  - Cup functions

    x^2, $|x|$, e^x
  - Cap functions

    $\log x$, $-x^2$
- Theorem 2.6.1: *If the function f has a second derivatiive which is no-negative (positive) everywhere, then the funtion is convex(strictly convex). (cup)*

Xi'an Jiaotong University

Jensen's inequality

- Theorem 2.6.2 (*Jensen's inequality*) *If f is a convex function and X is a random variable*,

  $$Ef(X)\geq f(EX)$$
- *Moreover, if f is strictly convex, the equality in* (*2.76*) *implies that X=EX with probability 1 (i.e. X is a constant)(cup)*
- The expectation of a cup function of a random variable is larger or equal to the cup function of the expectation of the random variable.

Xi'an Jiaotong University

Jensen's inequality

- If the function is a cap function, then we have: The expectation of a cap function of a random variable is less or equal to the cap function of the expectation of the random variable.
- Proof: Mathematical induction

For a two-mass-point distribution

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2),$$

Suppose that the theorem is true for distributions with k-1 mass points.

Jensen's inequality

$$\begin{aligned}\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} p_i' f(x_i) \\ &\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \\ &\geq f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p_i' x_i\right) \\ &= f\left(\sum_{i=1}^{k} p_i x_i\right),\end{aligned}$$

$$p_i' = p_i/(1-p_k)$$

Jensen's inequality

- Uniform distribution maximizes entropy
- Theorem 2.6.4 $H(X) \leq \log |\mathcal{X}|$, *where* $|\mathcal{X}|$ *denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over* $\mathcal{X}$.
- Proof: let $u(x)=1/|\mathcal{X}|$ be the uniform probability mass function over $\mathcal{X}$. And $p(x)$ be the probability mass function for X,Then

$$-\sum_x p(x)\log\frac{u(x)}{p(x)} \geq -\log\sum_x p(x)\frac{u(x)}{p(x)} = -\log\sum_x u(x) = 0$$

$$= \sum_x p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X) \qquad \rightarrow H(X) \leq \log|\mathcal{X}|$$

Jensen's inequality

$$I(X;Y) \geq 0$$

- Proof:

$$-I(X;Y) = \sum_x\sum_y p(x,y)\log\frac{p(x)p(y)}{p(x,y)}$$

$$\leq \log\sum_x\sum_y p(x,y)\frac{p(x)p(y)}{p(x,y)} = \log\sum_x\sum_y p(x)p(y) = 0$$

- Equation holds when $\frac{p(x)p(y)}{p(x,y)} = c \Rightarrow p(x,y) = p(x)p(y)$

The convexity of entropy

- The sum of cap(cup) functions is also a cap(cup) fucntion.
- The entropy of a random variable is a cap function
- Proof X $\quad P = \{p_1, p_2, \dots p_n\}$

$$H(X) = -\sum_i p_i \log p_i \qquad \text{Let } f(p) = -p\log p$$

Then $f'(p) = (-\log e)(\ln p + 1)$

$$f''(p) = -\log e \frac{1}{p} < 0$$

- $f(p)$ is a cup function
- $H(X)$ is also a cup function

The convexity of mutual information

- Theorem 2.7.4 *Let* $(X,Y) \sim p(x,y)$. *The mutual information* $I(X;Y)$ *is a cap function of* $p(x)$.

$$\overline{Q} = [Q(1),\dots Q(K)] \qquad I(X;Y) = \sum_{x,y} Q(x)p(y|x)\log\frac{p(y|x)}{\sum_z Q(z)p(y|z)}$$

- Proof: Let $I(X;Y) = f(\overline{Q})$
- Two distributions $\overline{Q}_1, \overline{Q}_2$ $\quad P_1(x,y) = Q_1(x)p(y|x) \quad P_2(x,y) = Q_2(x)p(y|x)$

$$P_1(y) = \sum_y P_1(x,y) \qquad P_2(y) = \sum_y P_2(x,y)$$

- Define a new distribution $\overline{Q} = \theta\overline{Q}_1 + (1-\theta)\overline{Q}_2$

$$P(x,y) = Q(x)p(y|x) = \theta P_1(x,y) + (1-\theta)P_2(x,y)$$

- We need to prove

$$\theta f(\overline{Q}_1) + (1-\theta) f(\overline{Q}_2) \leq f[\theta\overline{Q}_1 + (1-\theta)\overline{Q}_2]$$

The convexity of mutual information

$$\theta f(\bar{Q}_1)+(1-\theta)f(\bar{Q}_2)-f(\bar{Q})=\sum_{x,y}\theta P_1(x,y)\log\frac{p(y|x)}{P_1(y)}+\sum_{x,y}(1-\theta)P_2(x,y)\log\frac{p(y|x)}{P_2(y)}$$

$$-\sum_{x,y}[\theta P_1(x,y)+(1-\theta)P_2(x,y)]\log\frac{p(y|x)}{P(y)}$$

$$=\theta\sum_{x,y}P_1(x,y)\log\frac{P(y)}{P_1(y)}+(1-\theta)\sum_{x,y}P_2(x,y)\log\frac{P(y)}{P_2(y)}$$

$$\le 0$$

The convexity of mutual information

- Another proof
- $I(X;Y)=H(Y)-H(Y|X)=H(Y)-\Sigma p(x)H(Y|X=x)$
- Because $p(y|x)$ is fixed, $p(y)$ is a linear function of $p(x)$
- $H(Y)$ is a cup function of $p(y)$, is also a cap function of $p(x)$.
- $-H(Y|X)$ is a linear function of $p(x)$.
- Hence, $I(X;Y)$ is a cap function of $p(x)$.

Summary

- Entropy

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

- Source
- Joint entropy and conditional entropy

$$H(X,Y)=-\sum_x\sum_y p(x,y)\log p(x,y)$$

$$H(Y|X)=-E\log p(Y|X)$$

- Chain rule

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

Summary

- Channel
- Two circles
- Mutual information

$$I(X;Y)=\sum_x\sum_y p(x,y)\log\frac{p(y|x)}{p(y)}$$

- Some Inequalities
  - Jensen's inequality $Ef(X)\ge f(EX)$
  - H(X)$\le$log $|\mathcal{X}|$

    $$I(X;Y)\ge 0$$

  - $H(X)$ is a cap function of $p(x)$
  - $I(X;Y)$ is a cap function of $p(x)$