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

Independence Bound on Entropy

■ **From intuition to math expression**

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if the X_i are independent.

By chain rule for entropies,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i)$$

■ **Conditioning reduces entropy $H(Y|X) \leq H(Y)$**

■ Equality holds if and only if X_i is independent of $X_{i-1}, \ldots, X_1$ for all i (i.e., if and only if the X_i's are independent).

Intuition is not always correct

Markov Chain

Random variables $X,\ Y,\ Z$ are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

MC is a simple but very import structure for real world

- $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.
- $X \to Y \to Z$ implies that $Z \to Y \to X$. Thus, the condition is sometimes written $X \leftrightarrow Y \leftrightarrow Z$.
- If $Z = f(Y)$, then $X \to Y \to Z$.

$$I(X; Y|Z) = E_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}$$

If $X \to Y \to Z$, then $I(X; Z|Y) = 0$ (X and Z are conditionally independent given Y)

Data Processing Inequality

(Data processing inequality) If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$

Proof sketch: Expand $I(X;Y,Z)$ by chain rule

$$I(X;Y,Z) = I(X;Z) + I(X;\ Y|Z)$$
$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$$

where $\boldsymbol{I(X;Z|Y) = 0}$

- In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$.
- If $X \to Y \to Z$, then $I\ (X;Y|Z) \leq I(X;Y)$.
- Assume $X,\ Y$ are two independent random variables uniformly distributed on $\{0,1\}$.

$$Z = X + Y\ (\,mod\ 2)$$

  Calculate $I(X;Y|Z)$ $(I(X;Y|Z) > I(X;Y))$.

$I(X;Y;Z)$

- Assume X, Y are two independent random variables uniformly distributed on $\{0,1\}$.

$$Z = X + Y\ (\,mod\ 2)$$

Calculate $I(X;Y|Z)$ $(I(X;Y|Z) > I(X;Y))$.

Some facts:

- X, Y, Z are all uniformly distributed $H(X) = H(Y) = H(Z)$
- Any two of X, Y, Z can determine the other $H(X,Y,Z) = H(X,Y)$
- Any two of X, Y, Z are independent $H(X,Y) = H(X) + H(Y)$

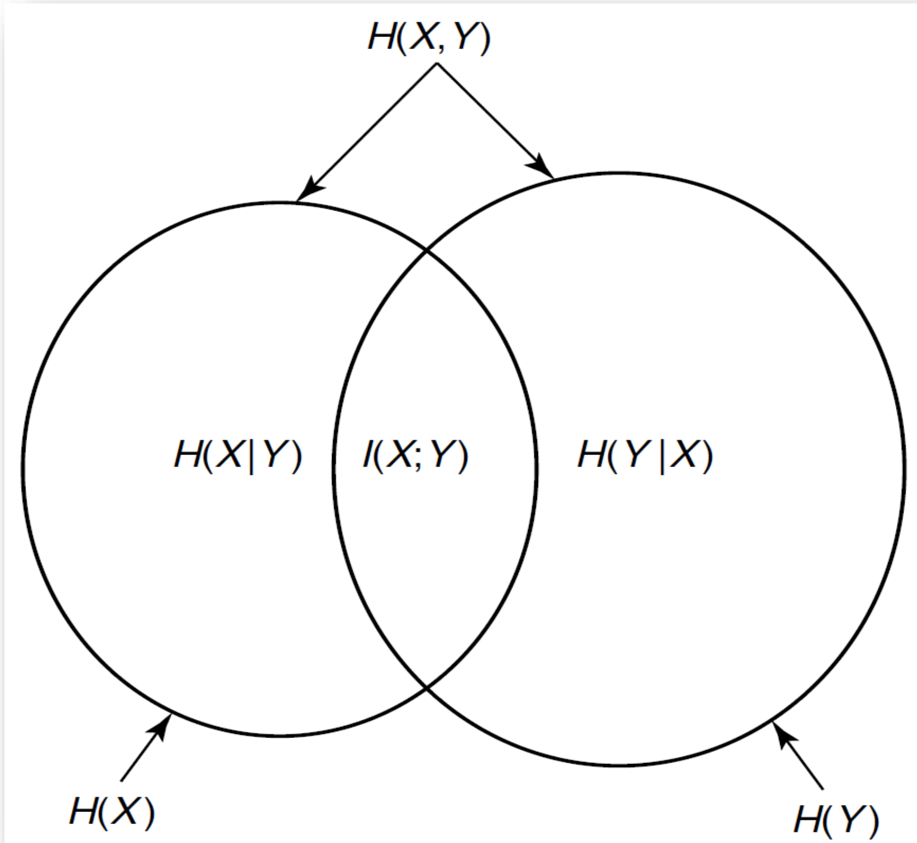
$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(X|Z) \\ &= H(X) \\ &= 1 \end{aligned}$$

$$I(X;Y|Z) > I(X;Y)$$

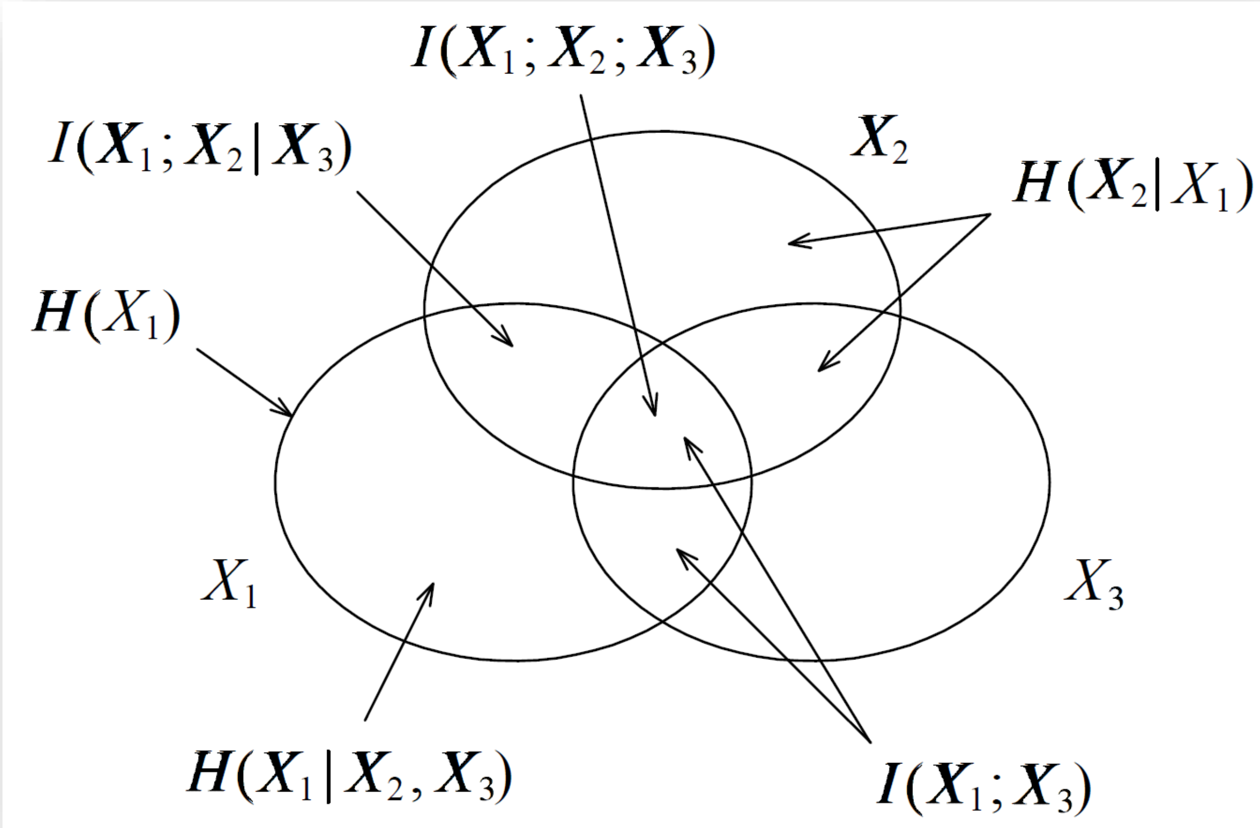
Define: $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$

Conditioning may not reduce mutual information. Mutual information is not uncertainty

Information Diagram: 2 RVs



Information Diagram: 3 RVs



- Area may be signed: negative
- Three circles: not three watches

Except $I(X_1;X_2;X_3)$, every part is ≥ 0. May be Negative!

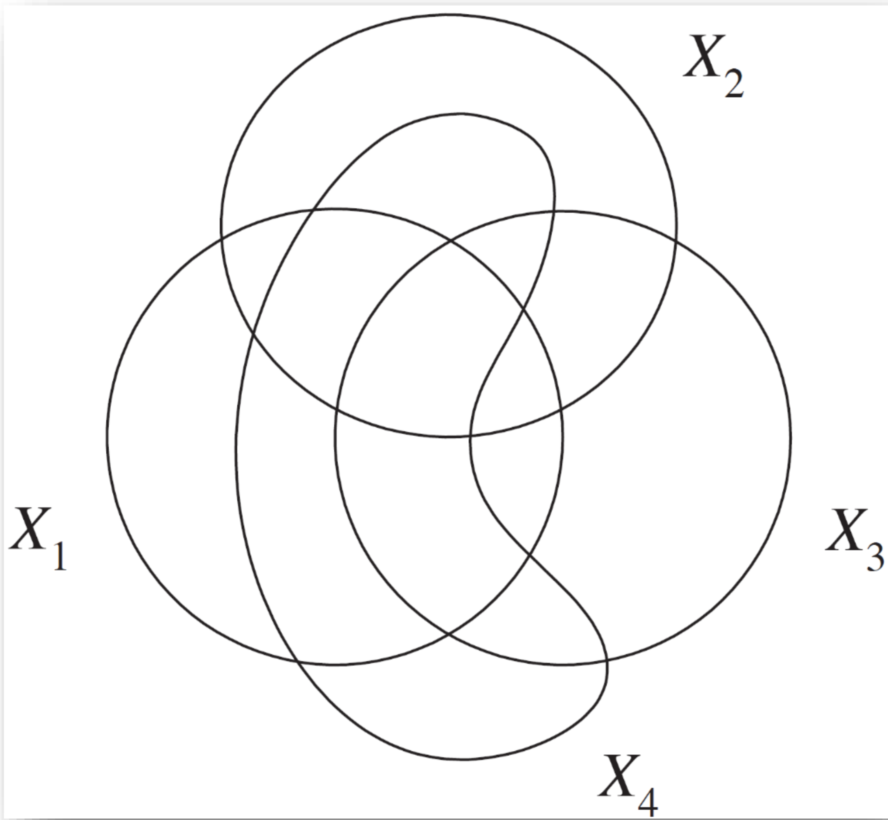
$Z = X + Y \ (mod\ 2)$

Reference: Ch. 3, Information Theory and Network Coding, R. W. Yeung

Information Diagram: 4 RVs

$H(X|Y)$

$I(X;Y|Z)$

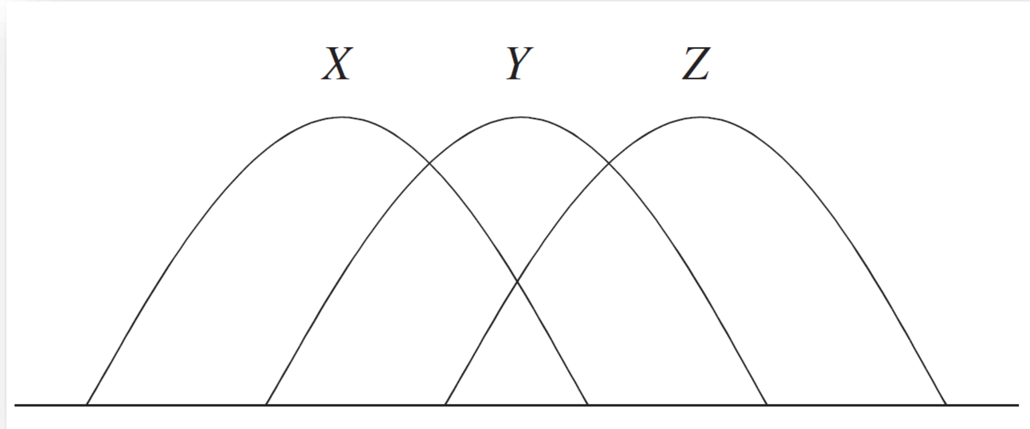


Only items like $I(X;Y|Z), H(X|Y) \geq 0$

Reference: Ch. 3, Information Theory and Network Coding, R. W. Yeung

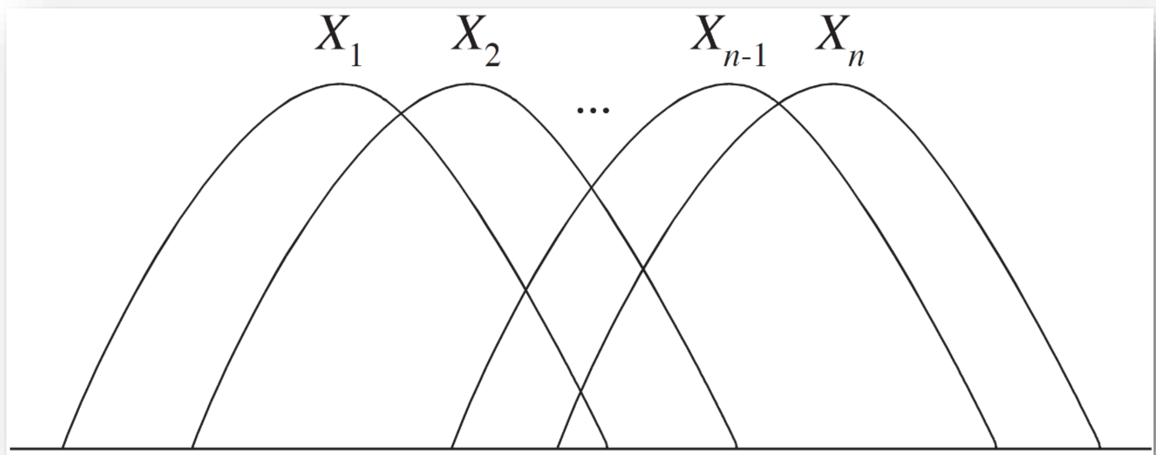
Information Diagram: Markov Chain

$X \rightarrow Y \rightarrow Z$



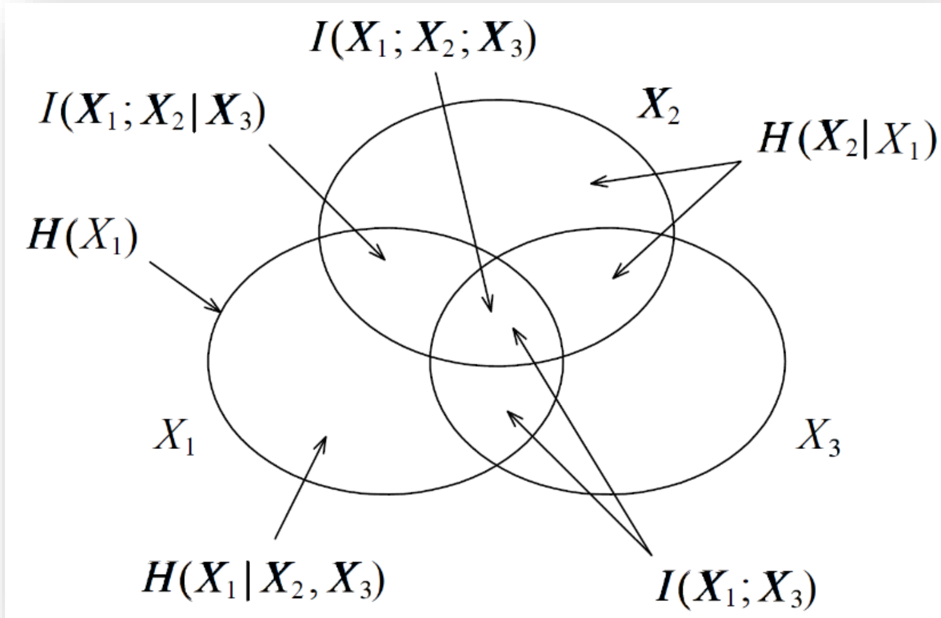
Each area ≥ 0

$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$



Reference: Ch. 3, Information Theory and Network Coding, R. W. Yeung

Examples



$$H(X,Y,Z) \leq \frac{H(X,Y)+H(Y,Z)+H(Z,X)}{2} \leq H(X)+H(Y)+H(Z)$$

$$H(X|Y,Z)+H(Y|X,Z)+H(Z|X,Y) \leq \frac{H(X,Y|Z)+H(Y,Z|X)+H(Z,X|Y)}{2} \leq H(X,Y,Z)$$

Examples (cont'd)

Homework 3

- Prove that under the constraint that $X \to Y \to Z$ forms a Markov chain, $X \perp Y|Z$ and $X \perp Z$ imply $X \perp Y$.
- Prove that the implication in (a) continues to be valid without the Markov chain constraint.
- Prove that $Y \perp Z|T$ implies $Y \perp Z|(X,T)$ conditioning on $X \to Y \to Z \to T$.
- Let $X \to Y \to Z \to T$ form a Markov chain. Determine which of the following inequalities always hold:

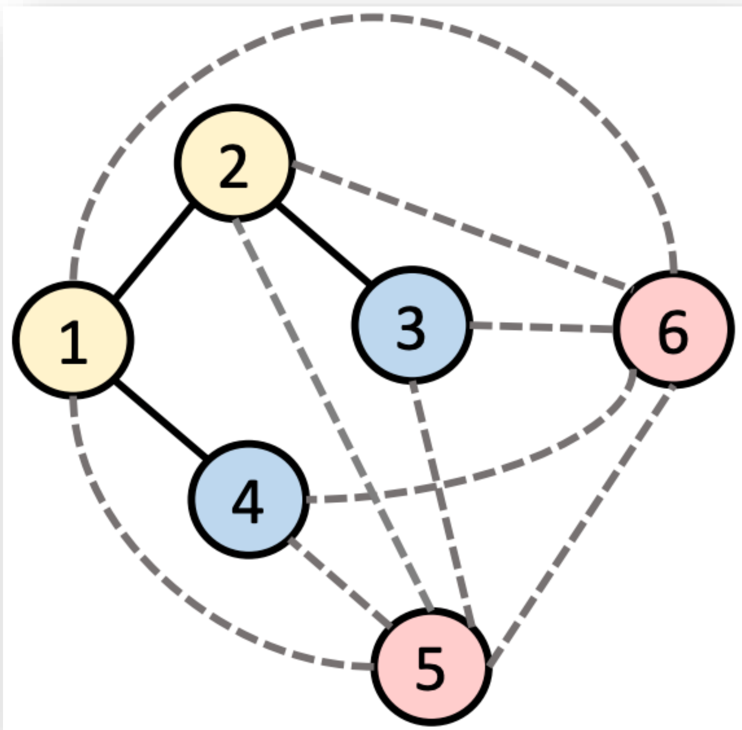
  $I.\quad I(X;T) + I(Y;Z) \geq I(X;Z) + I(Y;T)$

  $II.\quad I(X;T) + I(Y;Z) \geq I(X;Y) + I(Z;T)$

  $III.\quad I(X;Y) + I(Z;T) \geq I(X;Z) + I(Y;T)$

Example: Causality (因果推断)

给定条件：戴眼镜、爱好文学、弹吉他
推断：他/她是哪位同学

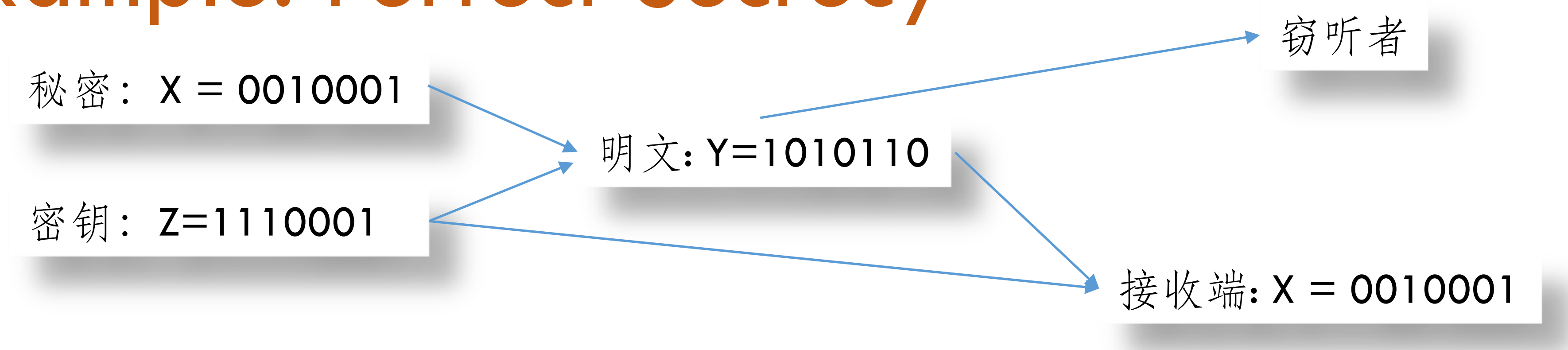


In information theory, we may use random variable to denote the conditions given in the problem, and apply the techniques in information measures to check whether a given condition is satisfied.

Given: $X \perp Y | Z$ and $X \perp Z$
Prove: $X \perp Y$

$$I(X;Y|Z) = 0, I(X;Z) = 0$$
$$I(X;Y) = 0$$

Example: Perfect Secrecy



Let X be the plain text, Y be the cipher text, and Z be the key in a secret key cryptosystem

- Y is generated from X and Z

$$H(Y|X,Z) = 0$$

- Since X can be recovered from Y and Z, we have

$$H(X|Y,Z) = 0$$

- We will show that this constraint implies

$$I(X;Y) \geq H(X) - H(Z)$$

- If the cipher text Y is required to be independent of the plain text X

$$I(X;Y) = 0$$

Then

$$H(X) \leq H(Z)\ (\text{信息长度小于密钥长度})$$

Fano's Inequality: Estimation

$X \rightarrow Y \rightarrow \hat{X} = g(Y)$

- Suppose that we wish to estimate a random variable X with a distribution $p(x)$.
- We observe a random variable Y that is related to X by the conditional distribution $p(y|x)$.
- From Y, we calculate a function $g(Y) = \hat{X}$, where $\hat{X}$ is an estimate of X and takes on values in $\hat{\mathcal{X}}$.
  - We will not restrict the alphabet $\hat{\mathcal{X}}$ to be equal to X, and we will also allow the function $g(Y)$ to be random.
- We wish to bound the probability that $\hat{X} \neq X$. We observe that $X \rightarrow Y \rightarrow \hat{X}$ forms a Markov chain. Define the probability of error

$$P_e = \Pr(\hat{X} \neq X)$$

- When $H(X|Y) = 0$, we know that $P_e = 0$. How about $H(X|Y)$, as $P_e \rightarrow 0$?

Fano: Establish the relation between P_e and $H(X|Y)$

Fano's Inequality

Theorem 2.10.1 (Fano's Inequality) For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr(\hat{X} \neq X)$, we have

$$H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

This inequality can be weakened to

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y) \quad \text{or} \quad P_e \geq \frac{H(X|Y)-1}{\log|\mathcal{X}|}$$

Intuition: $P_e \to 0$ implies $H(X|Y) \to 0$

Define an error random variable

$$E = \begin{cases} 0, & if\ \hat{X} = X \\ 1, & if\ \hat{X} \neq X \end{cases}$$

Then

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X,\hat{X}) \\ &= H(E|\hat{X}) + H(X|E,\hat{X}) \end{aligned}$$

Facts:

- $H(E|X,\hat{X}) = 0$
- $H(E|\hat{X}) \leq H(E) = H(P_e)$
- $H(X|E,\hat{X}) \leq P_e \log|\mathcal{X}|$

$$\begin{aligned} H(X|E,\hat{X}) &= \Pr(E=0)H(X|\hat{X}, E=0) + \Pr(E=1)H(X|\hat{X}, E=1) \\ &\leq (1-P_e)0 + P_e \log|\mathcal{X}|, \end{aligned}$$

Corollary. Let $P_e = \Pr(X \neq \hat{X})$, and let $\hat{X}: \mathcal{Y} \to \mathcal{X}$; then

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

Convexity/Concavity of Information Measures

(Log sum inequality) For nonnegative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i} = \text{const}$.

Prove via convexity/concavity

- (Concavity of entropy) $H(p)$ is a concave function of p.
- Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$.
  The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.
- (Convexity of relative entropy) $D(p||q)$ is convex in the pair (p, q); that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then
  $$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1 - \lambda)D(p_2||q_2)$$
  for all $0 \leq \lambda \leq 1$.

Homework 3:
Cover: 2.8, 2.9, 2.10 2.14, 2.15, 2.18, 2.20, 2.27, 2.32

Summary

The materials of this lecture are related to

- The textbook of T. Cover: 2.7, 2.8., 2.10
- The textbook of R. Yeung: 3.5, 3.6