# Lec1 -- Introduction to Information Theory

# Lec2 -- Probability Theory

Random Variable: $X$ Expectation: $EX$ Variance: $E(X-EX)^2$

Probability Distribution: Fuction $f$: $f(X)=P(X)$

# Lec3 -- Entropy

#### Shortest average encoding length -- Entropy

#### What

Encode $n$ events with probability $p_1, p_2, \ldots, p_n$ into $n$ different binary (0,1) strings.

Transfer messages for multiple times.

Goal: To minimize Message Length on average -- $E(l)=\Sigma p_i l_i$

Constraints: Decoded messages aren't ambiguous. -- sufficient condition(not necessary) - prefix-free codes

#### How

*Kraft Inequality* For prefix-free code $\Sigma 2^{-li} \leq 1$

Given r.v.(random variable) $X$,

pmf(probability math function) $(p_1, p_2, \ldots, p_n)$ $p_i \geq 0, \Sigma p_i = 1$

$min_{l_1, l_2, \ldots, l_n} \Sigma p_i l_i$ s.t. $\Sigma 2^{-li} \leq 1$ (the same as to be equal)

regard $q_i$ as $2^{-li}$

$min_{l_1, l_2, \ldots, l_n} \Sigma p_i log\frac{1}{q_i}$

$\Sigma p_i log\frac{p_i}{q_i} \geq 0$

$\Sigma p_i log\frac{1}{q_i} \geq \Sigma p_i log\frac{1}{p_i}$

#### Entropy

The lower bound of min code length on average: $\Sigma p_i log\frac{1}{p_i}$

Upper bound: $Entropy(p)+1$

$H(X)=\Sigma_i p_i log_2\frac{1}{p_i} \leq log_2 n$ -- Jensen's Inequality (convex)

#### Add property

r.v. $X-(p_1, p_2, \ldots, p_n)$ r.v. $Y-(q_1, q_2)$ r.v. $Z-(p_1, \ldots, p_{n-1}, q_1, q_2)$

$H(X)+p_n H(Y)=H(Z)$

**Optimal Code (Huffman Code)**

- Assume $p_i \geq p_j$, then $|c_i| \leq |c_j|$
- Kraft Inequality: $\Sigma_i 2^{-|c_i|} = 1$
- $|c_n| = |c_{n-1}|$
- $c_1, \dots, c_{n-1}^{'}$ is also an optimal code of $X$

# Lec4 -- Variations of Entropy

# Joint Entropy

r.v. $X, Y$ $P(X = x_i, Y = y_i)$ $i \in [m], j \in [n]$

$H(X, Y) := \Sigma_{i,j} p_{ij} log_2 \frac{1}{p_{ij}}$

$H(X) = \Sigma_i p_{x_i} log_2 \frac{1}{p_{x_i}}$

$H(Y) = \Sigma_i p_{y_i} log_2 \frac{1}{p_{y_i}}$

$H(X, Y) = H(X) + H(Y)$, if $X, Y$ independent else $\leq$

# Conditional Entropy

r.v. $X, Y$ $P(X = x_i, Y = y_i)$

Fix $x_i$ $P(Y|X = x_i)$

$H(Y|X = x_i) = \Sigma_j P(Y = y_j|X = x_i) log_2 \frac{1}{P(Y=y_j|X=x_i)}$

$H(Y|X) = \Sigma_i P(X = x_i) H(Y|X = x_i)$

- $H(Y|X) = H(Y)$, if $X, Y$ independent
- $H(Y|X) = 0$, if $X, Y$ fully dependent

$H(Y|X)$ represents the information of $Y$ given $X$.

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

# Mutual Entropy

Given Joint Entropy and Cond Entropy, there is:

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

Define:

$I(X; Y) := H(Y) - H(Y|X)$ equals $H(X) - H(X|Y)$

$I(X; Y) = \Sigma_{i,j} P(X = x_i, Y = y_j) log \frac{P(X=x_i, Y=y_j)}{P(X=x_i)P(Y=y_j)} \geq 0$

r.v. $X_1, X_2, \dots, X_m; Y_1, Y_2, \dots, Y_n$ $H(X_1^m, Y_1^n) = \Sigma P(X_1^m, Y_1^n) log \frac{1}{P(X_1^m, Y_1^n)}$

r.v. $X_1, X_2, \dots, X_m; Y_1, Y_2, \dots, Y_n$ $H(X|Y) = H(X, Y) - H(Y)$

r.v. $X_1, X_2, \dots, X_m; Y_1, Y_2, \dots, Y_n$

$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$

## Decomposition of Joint Entropy

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

## KL-divergence (Relative Entropy)

$P, Q$ are prob distributions $P = (p_1, \ldots, p_n), Q = (q_1, \ldots, q_n)$

$D(P||Q) := \Sigma_i p_i log \frac{p_i}{q_i} = \Sigma_i p_i log \frac{1}{q_i} - \Sigma_i p_i log \frac{1}{p_i}$

r.v. $X$, true $P$, estimated $Q$

$I(X;Y) = \Sigma_{x,y} P(X=x, Y=y) log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} = D(P(X,Y)||P(X)P(Y))$

$D(P||U_n) = log_2 n - H(P)$

**concav**

$P = (p_1, \ldots, p_n)$

$H(P) = H(p_1, \ldots, p_n)$

$H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q)$

$D(P||Q)$ given $P$, is $D$ convex ? given $Q$, is $D$ convex ? convexity of relative entropy

**convex**

*Convex:*

$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

*$\mu$-Strongly Convex:*$(\mu)$

$f(y) - f(x) \geq < \nabla f(x), y - x > + \frac{\mu}{2} ||y - x||^2$ holds for *any* $x, y$

**Thm (Pinsker's Ineq):**

$D(P||Q) \geq \frac{1}{2} ||P - Q||_1^2$

$plog\frac{p}{q} + (1-p)log\frac{1-p}{1-q} \geq 2(p-q)^2$

*Proof of Pinsker:*

$P = (p_1, \ldots, p_n)$ $Q = (q_1, \ldots, q_n)$

A $:= \{i : p_i \geq q_i\}$ B $:= \{i : p_i < q_i\}$

reduce $P, Q$ to Bernoulli distribution: $P', Q'$

$||P' - Q'||_1 = ||P - Q||_1$

$\Sigma_{i \in \mathrm{A}} p_i log \frac{p_i}{q_i} + \Sigma_{i \in \mathrm{B}} p_i log \frac{p_i}{q_i} = D(P||Q) \geq D(P'||Q') = \Sigma_{i \in \mathrm{A}} p_i log \frac{\Sigma p_i}{\Sigma q_i} + \Sigma_{i \in \mathrm{B}} p_i log \frac{\Sigma p_i}{\Sigma q_i}$

**Thm: Negative entropy is 1-strongly convex w.r.t. 1-norm:**

$\Sigma p_i log p_i - \Sigma q_i log q_i \geq < \nabla_p(\Sigma_i p_i log p_i), P - Q > + \frac{1}{2}||P-Q||_1^2$

**Hw**

1. Convexity of relative entropy
2. Pinsker's Ineq for Bernoulli distribution

## Data Processing Inequality

*No clever manipulation of the data can improve the inferences that can be made from the data.*

- R.v. $X, Y, Z$, Markov chain $X \to Y \to Z$;

  $P(Z|X,Y) = P(Z|Y) \leftrightarrow P(Z,X|Y) = P(Z|Y)P(X|Y)$

If $P(Z|X,Y) = P(Z|X)$, then $I(X;Y) \geq I(X;Z)$

$I(U;V,W) - I(U;V) = \Sigma_{u,v,w} P(u,v,w) log \frac{P(u,v,w)}{P(u)P(v,w)} - \Sigma_{u,v} P(u,v) log \frac{P(u,v)}{P(u)P(v)} = \Sigma_{u,v,w} P(u,v,w) log \frac{P(u,w|v)}{P(u|v)P(w|v)} \geq 0$

So $I(X;Y,Z) - I(X;Y) = 0$

$I(X;Y,Z) \geq I(X;Y)$

$\to I(X;Y) \geq I(X;Z)$

## Lec5 -- Entropy Rate

Regard Random Source $X_1, X_2, \ldots, X_t, \ldots$ as Stochastic Process $(X_t)_{t \geq 1}$

$E(l(X_1^T)) \in [H(X1, \ldots, X_T), H+1)$

- **Def 1**: The Entropy rate for a random source $X = (X_t)_{t \geq 1}$

  $H(X) = lim_{T \to \inf} \frac{1}{T} H(X_1, \ldots, X_T)$
- **Def 2**: $H(X) = lim_{T \to \inf} H(X_T | X_1^{T-1})$

  according to Entropy Decomposition

## Lec6 -- Differential Entropy

- **Def** : Differential Entropy

  Assume we have a conditional r.v. $X$ with density function $f(x)$

  $h(X) = -\int f(x) log(f(x)) dx$

$X$ discretization $\Delta \to discrete r.v. S_\Delta$

$h(X) = H(X_\Delta) - log \frac{1}{\Delta}$

*Discrete* r.v. $X, a > 0, b$

$H(X+b) = H(X) = H(aX)$

*Continuous* r.v.

$h(X+b) = h(X) \neq h(aX)$ **Hw1.**

- **Def** : Relative Entropy(KL-divergence)

$D(f||g) := \int f(x) log\frac{f(x)}{g(x)} dx$ where $f, g$ is density function

$D(f||g) = lim_{\Delta\to 0} D(P_\Delta||Q_\Delta)$

**Hw2. Is Entropy finite?**

# Lec7 -- Kolmogorov Complexity

## Kolmogorov Complexity

Entropy: minimum description length for random variables

What about deterministic object?

- **Def:** Kolmogrov Complexity

  The K-complexity for string $s$ w.r.t. Turing Machine $U$ is

  $K_U(s) := min_{U(p)=s}|p|$

- **Thm:** For any universal TM $u, u'$ and any $s \in \{0,1\}^*$

  $K_U(s) \le K_{U'}(s) + c$

**Hw.** Turing Machine, Universal TM, Computable, Halting Problem

- **Thm** : K-complexity is not computable

  **Proof** : Assume $\exists$ algorithm that computes K-complexity, so $\exists p$ finds the first string $s^*$ whose K-c $\ge 10^{10}$. Algorithm $p$ can be used to describe $s^*$.

## Maximum Entropy Principle

Estimate probability distribution of a r.v. $X$ -- $EX = \mu, VarX = \sigma^2$

**Hw.** MaxEnt Distribution -- $N(0, \sigma^2)$

*After Lec10*:

uniform distribution $u$

$0 \le D(f||u) = \int f \cdot ln\frac{f}{u} = -h(f) - \int flnu = -h(f) - \int ulnu = -h(f) + h(u)$

**Thm**: For random vector $X$, density function $EX = 0, Cov(X) = E[XX^T] = \Sigma$, $N(0, \Sigma)$ is the MaxEnt distribution.

**Prove**:

$0 \le D(f||g) = \int f \cdot ln\frac{f}{g} = -h(f) - \int flng = -h(f) - \int glng = -h(g) + h(u)$

$\int flng = \int f(x)[ln(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}) - \frac{1}{2}x^T\Sigma^{-1}x]dx$

where $\int g(x)x_ix_jdx = \int f(x)x_ix_jdx$

**Thm**: For random nonnegative integer $X$, $X = \mu$

$max_p \Sigma_{i\ge 0} p_i log\frac{1}{p_i} \;\; s.t. \;\; \Sigma_{i\ge 0} ip_i = \mu \;\; \Sigma_{i\ge 0} p_i = 1$

Lagrange: $p_k \; \propto e^{-ck}$

**Hw.** Exp distribution is MaxEnt.

**Thm**: Concave

$log\ det(\Sigma)$ is a concave function.

**Prove**: fix $\Sigma_0$, $g(t) = log|\Sigma_0 + t\Sigma|$

if $g(t)$ is concave for $t \in [0, 1]$, then $log\ det$ is concave.

As $\Sigma_0$ is positive definite, we can decompose $\Sigma_0 = QQ^T$

reduce to $g(t) = log|I + tV|$, decompose $V = PAP^T$ where $P$ is orthonormal matrix, elements in $A$ are eigenvalue.

reduce to $g(t) = log|I + tA|$

**Prove**:

$\Sigma_1, \Sigma_2$ p.d. $\lambda \in [0, 1]$

$logdet(\lambda\Sigma_1 + (1-\lambda)\Sigma_2) \geq \lambda logdet(\Sigma_1) + (1-\lambda)logdet(\Sigma_2)$

Constuct $X_1, X_2$ r.v., $X_1 \sim N(0, \Sigma_1), X_2 \sim N(0, \Sigma_2)$

r.v. $K$, $P(K = 1) = \lambda, P(K = 0) = 1 - \lambda$

$Z = X_1\ if\ K = 1\ else\ X_2$

So: $Cov(Z) = \lambda\Sigma_1 + (1-\lambda)\Sigma_2$

$h(Z) \leq \frac{1}{2}log[(2\pi e)^n|\lambda\Sigma_1 + (1-\lambda)\Sigma_2|]$ (MaxEnt of Gaussian Distribution)

$h(Z) \geq h(Z|K) = \lambda h(Z|K = 1) + (1-\lambda)h(Z|K = 0) = \lambda h(X_1) + (1-\lambda)h(X_2)$

Q.E.D.

$X \sim N(0, \Sigma)$

$h(X) = \frac{1}{2}log((2\pi e)^n|\Sigma|)$ bits *Compute trick:* $tr(AB) = tr(BA)$

# Lec8 -- Channel Coding: Algorithms

*Map string* $\{0, 1\}^m$ to $\{0, 1\}^n$ *with maximum Hamming Distance*

$\{0, 1\}^m \to \{0, 1\}^n$

$N = 2^n, M = 2^m, V_B = \Sigma_{k=0}^{r/2}\binom{n}{k}$

- **Thm** : Chernoff Bound

  iid. Bernoulli r.v. $X, X_1, \dots, X_n,\ \ EX = p$

  $P(\frac{1}{n}\Sigma_i X_i \geq p + \delta) \leq 2^{-nD_B(p+\delta||p)} \quad where \quad D_B(p+\delta||p) = (p+\delta)log_2\frac{p+\delta}{p} + (1-p-\delta)log_2\frac{1-p-\delta}{1-p}$

  **Proof** :

  1) Chernoff Ineq : r.v. $Y$ $P(Y \geq k) = P(e^{tY} \geq e^{tk})$

  Markov Ineq: $\leq inf_{t>0}Ee^{tY}e^{-tk}$

  2) $P(\frac{1}{n}\Sigma X_i \geq p + \delta) = P(\Sigma X_i \geq n(p+\delta)) \leq inf_{t>0}Ee^{t\Sigma X_i}e^{-nt(p+\delta)}$

$X_i\ \ iid \to Ee^{t\Sigma X_i} = (Ee^{tX})^n = [pe^t + 1 - p]^n$

**Hw.** Chernoff Bound

- **Thm** : Gilert-Vashamov Bound

  If $n \geq \frac{2m}{1-H(\delta)}$ where $H(\delta) = -(\delta log\delta + (1-\delta)log(1-\delta))$ $\delta \in (0, \frac{1}{2})$

  then there exists $c_1, c_2, \ldots, c_{2^m} \in \{0,1\}^n$

  such that $d_H(c_i, c_j) \geq \delta n$

  **Proof** : Probabilistic Method

  Uniformly random chooses two strings $\in \{0,1\}^n, S, S'$

  $P(d_H(S, S') \leq \delta n) \leq 2^{-n(1-H(\delta))}$ // Chernoff Bound

  Uniformly random chooses $2^m$ strings $\in \{0,1\}^n$

  $P(\exists_{i\neq j} i, j \in [2^m], d_H(c_i, c_j) \leq \delta n) \leq 2^{2m} 2^{-n[1-H(\delta)]}$

  When $n \geq \frac{2m}{1-H(\delta)}$, $P < 1$ Q.E.D.

Decoding: find the nearest neighbor of encoded message.

Codes should share some special structure to design efficient decode algorithm.

- **Hamming Codes(7,4)** : 1) $d_H(c_i, c_j) \geq 3bits$; 2) Coding/Decoding Computationally efficient

  $GF(2)$

  The kernel space (null space) of $H$ $dim(ker(H)) = 7 - 3 = 4$

  $|ker(H)| = 2^4 = 16$ $c, c' \in ker(H)$ $d_H(c, c') = |c_i + c_j|_1 \geq 3$ $c_i + c_j \in ker(H) - \{0^7\}$

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{H}$$

  - Encoding: $\{0,1\}^4 \to \{0,1\}^7$ $\{0,1\}^4 \to ker(H)$
  - Decoding: $\{0,1\}^7 \to ker(H)$

    $HS = H(c + e_i) = He_i$
  - Encoding: $\{0,1\}^4 \to ker(H)$ $H = [P_{3*4}\ \ I_{3*3}]$ $G = [I_{4*4}\ \ -P^T]$

    $HG^T = 0$ $c = mG \in ker(H)$

# Lec9 -- Communication Complexity

# Deterministic Algorithm

Setting: Alice and Bob compute $f(x,y), x, y \in \{0,1\}^n$

$x \in$ Alice $y \in$ Bob

Communication: # of bits communicated

1) protocol design (UB)

2) hardness (LB)

$f(x,y) = 1\ if\ x = y\ \ else\ \ 0$ $CC(f_{Eq}) \geq \Omega(n)$ --Deterministic protocol matrix $2^n * 2^n$ $x(f)$: minimum number of chronomic rectangles

- **Thm1:** $CC(f) \geq log_2 \chi(f)$

1) Lower bound $\Omega(log_2 \chi(f))$

2) Upper bound in terms of $\chi(f)$?

- **Thm2:** $log_2 \chi(f) \leq CC(f) \leq O(log_2^2 \chi(f))$

**Proof:** Represent each rectangle w/ $log_2 \chi(f)$ bits

Define:

- For a rectangle $R$, define $K_x(R) = \#$ of rectangles have overlap w/ $R$ in rows.
- For a rectangle $R$, define $K_y(R) = \#$ of rectangles have overlap w/ $R$ in columns.

Protocol: For $t = 1, 2, \ldots$

1. Alice choose a rectangle $R$ such that $x \in R$, and $K_x(R)$ is the smallest among all rectangles still active. Remove all rectangle not overlap w/ $R$ in rows($M_1$).
2. Bob choose a rectangle $R'$ such that $y \in R'$, and $K_y(R')$ is the smallest among all rectangles still active. Remove all rectangle not overlap w/ $R'$ in columns($M_2$).

$M_1 = N - K_x(R)$ $M_2 = N - K_y(R')$

$K_x(R) + K_y(R) < N$ $K_x(R) \leq K_x(R')$

so. $max(M_1, M_2) \geq \frac{N}{2}$

$Rank(M) \leq \chi(f)$ Matrix decomposition -- $Rank(A + B) \leq Rank(A) + Rank(B)$

- **Log rank Conjecture**

  $CC(f) \leq polylog(Rank(M_f))$

- **Best Upper Bound**

  $CC(f) \leq \frac{Rank(M_f)}{log(Rank(M_f))}$

  $CC(f) \leq \sqrt{Rank(M_f)}$

**Hw:**

1)

$f(x, y) = < x, y > = \bigoplus x_i y_i$

$g(x, y) = (-1)^{f(x,y)}$ matrix $g$ is orthogonal.

**Walsh-Hadamad matrix**

2)

Graph $G = < V, E >$

Alice has a clique $C \subseteq G$, Bob has an independent set $I \subseteq V$

Goal: Decide if $C \; interset \; I = \emptyset$

Design a protocol: as small number of bits as possible in terms of $n = |V|$.

Optional: lower bound

## Randomized Algorithm

Consider $f(x, y) = 1_{x==y}$

$P \in [n^2, n^3]$ Polynomial on $Z_P$

Protocol:

1) Alice uniformly randomly select a $t \in \{0, 1, \dots, p-1\}$ and construct polynomial

$u = x_{n-1}t^{n-1} + x_{n-2}t^{n-2} + \dots + x_0 (mod\, p)$, send $u, t$ to Bob.

2) Bob calculate $v$ and check if $u == v$.

$Rootnumber \le n-1, ErrorProb \le \frac{n-1}{p}$

**Extention**

Multi-agent communication: one send message to everyone else.

$f(x, y, z) = MajorityFunction\ x, y, z \in \{0, 1\}^n\ f \in \{0, 1\}$

$f_{MJ}(x, y, z) = \bigoplus_{i=1}^n Majorityvote(x_i, y_i, z_i)$

**Hw1.** Majorityvote $CC(f)$

**Hw2.** Number on your forehead setting

## Lec10 -- Fisher Information

### Fisher Information and Cramer-Rao Inequality

- Sample $X = (X_1, \cdots, X_n)$ (typically X_1,···, X_n iid)

  $f(x; \theta) = \Pi_{i=1}^n f(x_i; \theta)$ -- probability density function

  Estimator: $\Phi : X \to \theta$

  - unbiased: $E[\Phi(X)] = \theta$
  - the lower bound of variance: $Var(\Phi(X))$
- **Def:** (Score function) For a sample $X = (X_1, \cdots, X_n)$, let $f(x; \theta)$ be the density function of the sample. The score function is defined as:

  $S(X; \theta) = \frac{\partial}{\partial \theta} ln(f(X; \theta))$

  $E(S(X; \theta)) = \int S(X; \theta) f(X; \theta) dx = \int \frac{\partial}{\partial \theta} f(X; \theta) dx = \frac{\partial}{\partial \theta} \int f(X; \theta) dx = 0$
- **Def:** (Fisher Information) The Fisher Information of $\theta$ w.r.t. sample $X$ is defined as $I(\theta) := E[S(X; \theta)^2] = \int (\frac{\partial}{\partial \theta} ln f(X; \theta))^2 dx$
- **Proposition:** $I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} ln f(X; \theta)]$

  Proof: $E[\frac{\partial^2}{\partial \theta^2} ln f(X; \theta)] = \int \frac{\partial^2}{\partial \theta^2} ln f(X; \theta) f(X; \theta) dx = \int [\frac{-(\frac{\partial}{\partial \theta} f(X; \theta))^2}{f^2(X; \theta)} + \frac{\frac{\partial^2}{\partial \theta^2} f(X; \theta)}{f(X; \theta)}] f(X; \theta) dx = -E(S(X; \theta)^2)$
- **Thm(Cramer-Rao Inequality)**

  For any unbiased estimator $\Phi : X \to R$, $Var(\Phi(X)) \ge \frac{1}{I(\theta)}$

  Proof: $I(\theta) = Var(S(X; \theta)) = E[S^2(X; \theta)]$

Cauchy Inequality: $Var(\Phi(X))Var(S(X;\theta)) \geq E[(\Phi(X)-E\Phi(X))(S(X;\theta)-ES(X;\theta))]^2 = E[\Phi(X)S(X;\theta)]^2$

$E[\Phi(X)S(X;\theta)] = \int \Phi(X)\frac{\partial}{\partial\theta}lnf(X;\theta)f(X;\theta)dx = \frac{\partial}{\partial\theta}E(\Phi(X)) = 1$

### Fisher Information for Multiple Parameters

Sample vector $X$

$\hat{\theta} = \phi(X)$ $\hat{\theta} \in R^k$ Estimate $Cov(\phi(X))$

$I(\theta) = E[\nabla_\theta ln(f(X;\theta)\nabla_\theta lnf(X;\theta)^T] = Cov(S(X;\theta))$

- **Thm(Cramer-Rao Inequality)**

  Every unbiased esitimator $\phi$ satisfies:

  $Cov(\phi(X)) \succeq I(\theta)^{-1}$ $A \succeq B$ means $A-B$ is a positive definite matrix.
- A simplified version: Estimate $q(\theta_1,\ldots,\theta_k), q: R^k \to R$, if $\phi$ is an unbiased estimator or $q(\theta)$, then:

  $Var(\phi(X)) \geq \nabla_\theta q(\theta)^T I(\theta)^{-1}\nabla_\theta q(\theta)$

$E[\phi(X)] = q(\theta)$

$\nabla_\theta q(\theta) = \nabla_\theta \int \phi(X)f(X;\theta)dx = \int \phi(X)\frac{\nabla_\theta f(X;\theta)}{f(X;\theta)}f(X;\theta)dx = E[\phi(X)S(X;\theta)] = E[(\phi(X)-E[\phi(X)])S(X;\theta)]$

Then we have:

$\nabla_\theta q(\theta)^T I(\theta)^{-1}\nabla_\theta q(\theta) = \nabla_\theta q(\theta)^T I(\theta)^{-1}S(X;\theta)E[\phi(X)-E[\phi(X)]]$

Cauchy Inequality:

$\leq Var(\phi(X))^{1/2}\{\nabla_\theta q(\theta)^T I(\theta)^{-1}S(X;\theta)S(X;\theta)^T I(\theta)^{-1}\nabla_\theta q(\theta))\}^{1/2} = Var(\phi(X))^{1/2}\{\nabla_\theta q(\theta)^T I(\theta)^{-1}\nabla_\theta q(\theta)\}^{1/2}$

**Thm: (Fano's Inequality)**

Send message $X$, receive message $Y$.

$P_e \geq \frac{H(X|Y)-1}{log|H|}$ $H$ is the support

Proof:

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

$H(X|Y, X \neq g(Y)) \leq log|H|$

Define r.v. $E$ as $E = 0\ if\ X = g(Y)\ else\ 1$

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

$H(X,E|Y) = H(E|Y) + H(X|E,Y) \leq 1 + P(E=0)H(X|E=0,Y) + P(E=1)H(X|E=1,Y) \leq 1 + P_e log|H|$

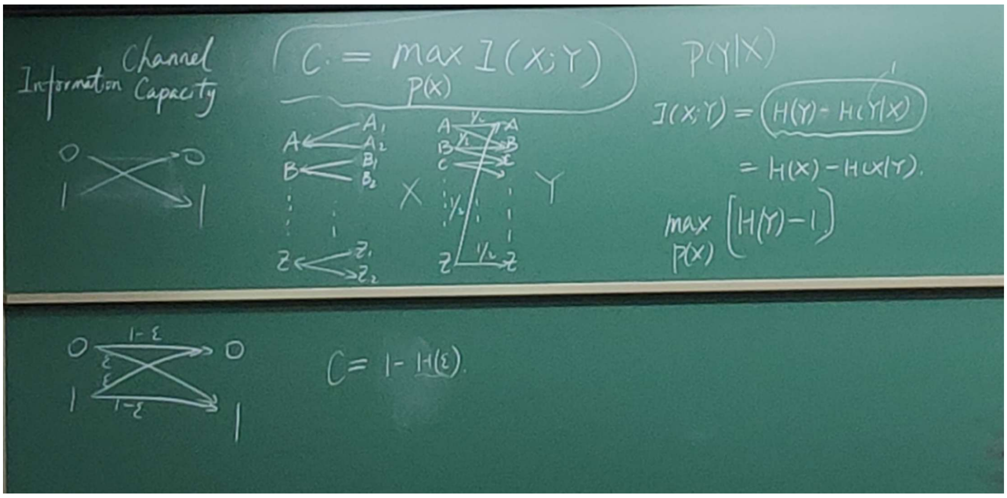
Q.E.D.

## Lec11 -- Channel Capacity

1. Implementation (algorithm)

   Encoding/Decoding constraint: $err \to 0$, efficiency
2. Conceptual

- **Def**: (Channel Capacity) := $max_{P(X)} I(X;Y)$




**AEP**: Asymptotic Equipartition property

**The Law of Large Number**:

$P(|\frac{1}{n}\Sigma_{i=1}^{n} X_i - EX| \geq \epsilon) \to 0$

if $g(X)$ subject to some property(e.g. like random variable):

$P(|\frac{1}{n}\Sigma_{i=1}^{n} g(X_i) - Eg(X)| \geq \epsilon) \to 0$

$g(X) = -log\, p(X)$

$P(2^{-n(H(x)+\epsilon)} \leq P(X_1, X_2, \ldots, X_n) \leq 2^{-n(H(x)-\epsilon)}) \to 1$

$P(X_1, \ldots, X_n) \approx 2^{-nH(X)}$ with high probability

**Typical Sequence & Set**:

$X_1, \ldots, X_n$ is a typical sequence if $P(X_1, \ldots, X_n) \in 2^{-n[H(X) \pm \epsilon]}$

Typical set = {typical sequence}

$P(X_1, \ldots, X_n) \approx 2^{-nH(X)}$

So $|typical\ set| \approx 2^{nH(X)}$

So we can assume that all sequences are uniformly distributed in typical set.

**Jointly Typical Sequence & Set**

$(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$

// $P(|\frac{1}{n}\Sigma_{i=1}^{n} - logP(X_i, Y_i) - H(X, Y)| \geq \epsilon) \to 0$

$| - \frac{1}{n}\Sigma_i logP(X_1, \ldots, X_n; Y_1, \ldots, Y_n) - H(X;Y)| \leq \epsilon$

$| - \frac{1}{n}\Sigma_i logP(X_1, \ldots, X_n) - H(X)| \leq \epsilon$

$| - \frac{1}{n}\Sigma_i logP(Y_1, \ldots, Y_n) - H(Y)| \leq \epsilon$

Jointly Typical Sequence.

1) $P(X_1, Y_1, \ldots, X_n, Y_n) \approx 2^{-nH(X,Y)}$

2) $P(X_1, \ldots, X_n) \approx 2^{-nH(X)}$

3) $P(Y_1, \ldots, Y_n) \approx 2^{-nH(Y)}$

*Random draw of jointly typical sequence*

1) $(x_i, y_i) \sim P(X, Y)$

2) $x_i \sim P(X), y_i \sim P(Y|X_i)$

Q: On average, for each typical sequence $(X_i, \ldots, X_n)$, # of $(X_i, Y_i, \ldots, X_n, Y_n)$ with sequence $X$ is $2^{nH(Y|X)}$.

*Random draw*

$X_1, \ldots, X_n \sim P(X)$ $Y_1, \ldots, Y_n \sim P(Y)$

Q: On average, the probability that $(X_1, Y_1, \ldots, X_n, Y_n)$ is a jointly typical sequence is $2^{-nI(X,Y)}$.

**Setting**

Channel $P(Y|X)$

Input $X_1, X_2, \ldots, X_n$ iid discrete

Input $Y_1, Y_2, \ldots, Y_n$ iid discrete

$W = \{1, 2, \ldots . n\}$ message (uniform)

Coding: $W \to \mathcal{X}^n$ $log\, m/n$ bits per trans

Decoding: $g :\ Y^{(n)} \to W$

Error rate: $Pr[g(Y^n) \neq w | X_n = X_n(w)]$

$M = 2^{nR}$ keeps efficiency $R$.

If $lim_{n\to\infty} \lambda_{max} = 0$, $\lambda_{max} = max_{i \in [2^{nR}]} \lambda_i$

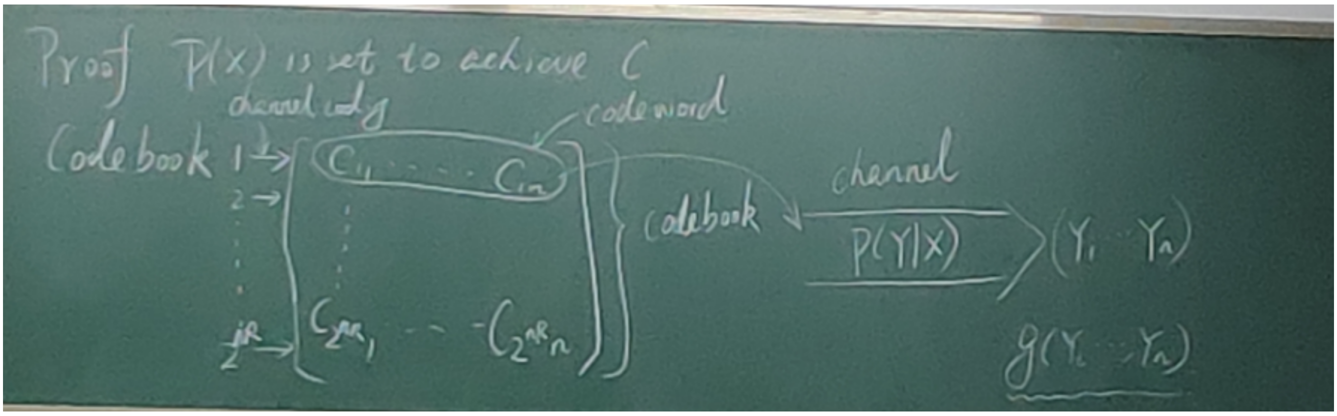
**Thm (Channel Coding Thm)**

$C = max_{P(X)} I(X; Y)$

If $R < C$, then, $\exists$ a sequence of $(2^{nR}, n)$ codes, such that $lim_{n\to\infty} \lambda_{max}^{(n)} = 0$.

If $R > C$, then there is no coding method such that $lim_{n\to\infty} \lambda_{max}^{(n)} = 0$.

$P(X)$ is set to achieve $C$.




1) Random endoding

$c_{ij} \sim P(X)$ iid for all $i, j$

2) Decoding

On receiving $Y_1, \dots, Y_n$, if there exists a unique codeword $c_{i1}, \dots, c_{in}$, such that $(c_{i1}, \dots, c_{in}; Y_1, \dots, Y_n)$ is a jointly typical sequence, then decode $g(Y_1, \dots, Y_n)$ as $i$, else report a failure.

*Error probability*

- Y is not a typical sequence. -- low
- X,Y is not a jointly typical sequence. -- low
- $\exists X' s.t. (X', Y)$ is jointly typical sequence. -- $2^{-nI(X;Y)}$

3) From average err. to max err.

**Proof**

- Part I

$R < C$, $P(X) = argmax_{P(X)} I(X;Y)$

Average err over all codebooks and messages:

$P(Err) = \Sigma_{CB} P(CB) \frac{1}{2^{nR}} \Sigma_{i=1}^{nR} P_e^{CB}(w_i) = \frac{1}{2^{nR}} \Sigma_{i=1}^{nR} \Sigma_{CB} P(CB) P_e^{CB}(w_i)$

$P(Err) = \Sigma_{CB} P(CB) P_e^{CB}(w_i) \leq \epsilon + \epsilon + 2^{-nI(X;Y)} * 2^{nR} = \epsilon' + 2^{-n(C-R)}$

Therefore, there exists a CodeBook such that error prob over all messages is small.

For any message, consider the best half CodeBooks $2^{nR-1}$, there is $max_{i \in [2^{nR-1}]} \lambda_i \leq 2\epsilon$.

$(2^{n(R-\frac{1}{n})}, n)$

- Part II

Fano's Inequality

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

$R > C$, r.v. $W \in_R 1, 2, \dots, 2^{nR}$ $X -> Y$

$nR = H(W) = H(W|Y_1, \dots, Y_n) + I(W; Y_1, \dots, Y_n)$

$nR \leq P_e^{(n)} nR + 1 + I(X_1^n; Y_1^n) = P_e^{(n)} nR + 1 + nC$

$P_e^{(n)}$ cann't goes to 0.

## Lec12 -- Rate Distortion Theory

Quantification:

$X \sim U\{a, b, c\}$ $H(X) = log_2 3 bits$

$d(x, x') = I[x \neq x']$

$D = \Sigma_x p(x) d(x, \phi(x)) = 1/3$

$\phi^{(n)} := \{a, b, c\}^n -> \{a, b, c\}^n$ $|\phi^{(n)}| = 2^n$

**Def:** $\phi^{(n)}$ is a mapping: $\phi^{(n)} := X^n -> X^n$. Say $\phi^{(n)}$ is a $(2^{nR}, n)$ rate distribution code if $|\phi^{(n)}| \leq 2^{nR}$.

**Def:** $D := E[d(X^n, \phi(X^n))]$

Given $D$, find the best encoding method to minimize $R$.

**Thm:** $R^{(I)}(D) = min_{P(X'|X)} I(X; X')$ $E[d(X; X')] \leq D$