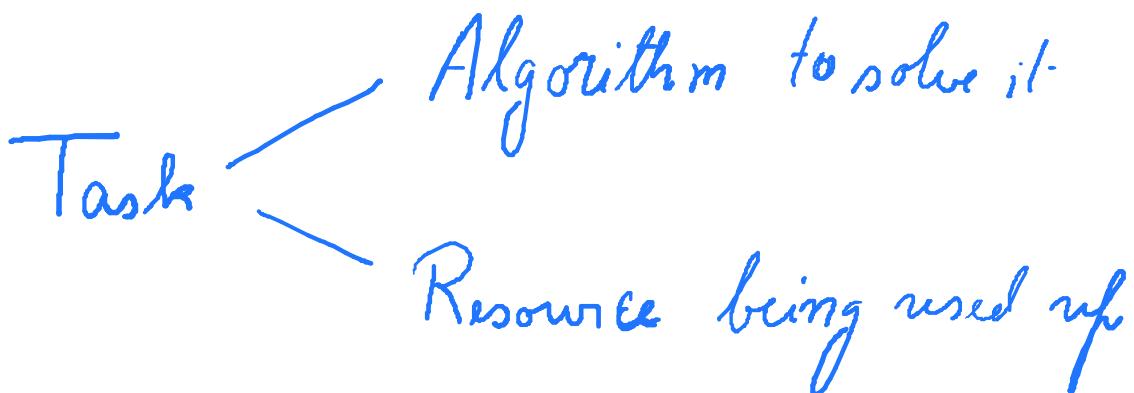


* Background



Qstn:- What is the least possible resource usage among all possible algorithms that solve this task?

- Information Theory is one set of tools and techniques developed to tackle such questions.

In this course, we shall tackle three such tasks

- i) Lossless source coding
- ii) Lossy source coding

iii) Channel coding

* Some mathematics for warm up

i) Metric spaces

A set X and a map $d: X \times X \rightarrow \mathbb{R}_{\geq 0}$ constitute a metric space if

- i) $d(x, y) \geq 0$ & $d(x, y) = 0$ iff $x = y$
- ii) $d(x, y) = d(y, x)$
- iii) $d(x, y) \leq d(x, z) + d(y, z)$

e.g. $(\mathbb{R}, |\cdot|)$, $(\mathbb{R}^2, \|\cdot\|_2)$ etc.

→ Total variation distance

a) Let $\mathcal{P}_X = \{P: P \text{ is a pmf of } X\}$

Define, $TV: \mathcal{P}_X \times \mathcal{P}_X \rightarrow \mathbb{R}_{\geq 0}$

as

$$TV(P, Q) := \frac{1}{2} \sum_{x \in X} |P(x) - Q(x)|$$

Then, (P_X, TV) is a metric space. To see this

$TV(P, Q) \geq 0$ obvious.

Also, as $|P(x) - Q(x)| \geq 0$

$TV(P, Q) = 0 \Rightarrow P(x) = Q(x) \quad \forall x \in X$

Thus $P = Q$.

$TV(P, Q) = TV(Q, P)$ is obvious.

$TV(P, R) + TV(Q, R)$

$$= \frac{1}{2} \sum_{x \in X} (|P(x) - R(x)| + |R(x) - Q(x)|)$$

$$\geq \frac{1}{2} \sum_{x \in X} |P(x) - Q(x)| \quad [\Delta \text{ inequality}]$$

$$= TV(P, Q)$$

b) Let $\underline{P_R} = \{f: f \text{ is a pdf}\}$

Define
$$TV(f, g) = \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| dx$$

Again, it is easy to see that

$(\underline{P_R}, TV)$ is a metric space.

Note :- The above two definitions of TV are special cases of the general definition of TV. There are in fact other distances between probability measures such as Wasserstein, Levy, Levy-Prokhorov etc.

ii) Limits of sequences

Let (X, d) be a metric space.

Let $a_{n,n \geq 1}$ be a sequence from X .

We say that $a_n \rightarrow a$ if for

every $\varepsilon > 0$, $\exists N(\varepsilon)$, s.t. $\forall n \geq N$,

$$d(a_n, a) \leq \varepsilon.$$

e.g. 1. $(\mathbb{R}, |\cdot|)$, $a_n = \frac{1}{n}$. $a = 0$.

Then, fix any $\varepsilon > 0$.

$$|a_n - a| = \left| \frac{1}{n} \right| \leq \varepsilon, \quad \forall n \geq \left\lceil \frac{1}{\varepsilon} \right\rceil$$

2. (\mathbb{R}_+, TV) .

$$a_n = \text{unif} [0, 1 + \frac{1}{n}]$$

$$a = \text{unif} [0, 1]$$

Fix any $\varepsilon > 0$.

$$TV(a_n, a) = \frac{1}{2} \left[\int_0^1 \left(1 - \frac{1}{1 + \frac{1}{n}} \right) \cdot dx + \int_1^{1 + \frac{1}{n}} \frac{1}{1 + \frac{1}{n}} \cdot dx \right]$$

$$= \frac{1}{2} \left(1 - \frac{1}{1 + \frac{1}{n}} + \frac{\frac{1}{n}}{1 + \frac{1}{n}} \right)$$

$$= \frac{\frac{1}{n}}{1 + \frac{1}{n}}$$

$$= \frac{1}{n+1}$$

Now $\forall n \geq \lceil \frac{1}{\varepsilon} - 1 \rceil$

$$\frac{1}{n+1} \leq \varepsilon.$$

Thus, $\text{TV}(a_n, a) \leq \varepsilon$.

Thus, $a_n \rightarrow a$.

iii) Convergence in probability and the weak law of large numbers

Let X_1, X_2, \dots be a sequence of

random variables.

We say that $X_n \xrightarrow{p} X$ if for any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \rightarrow 0.$$

Theorem (Weak Law of Large numbers):-

Let $X_n, n \geq 1$ be a sequence of independent and identically distributed random variables. Then,

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} E[X].$$

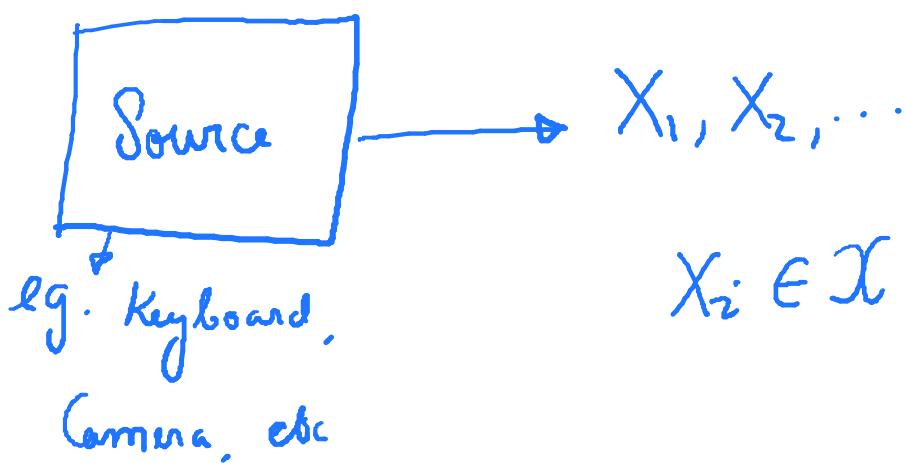
* Lossless Source Coding

- Key idea

To compress information for storage such that the entire information can be recovered.

e.g. ZIP files, TIFF files

Setting :-



\mathcal{X} is called the alphabet of the source.

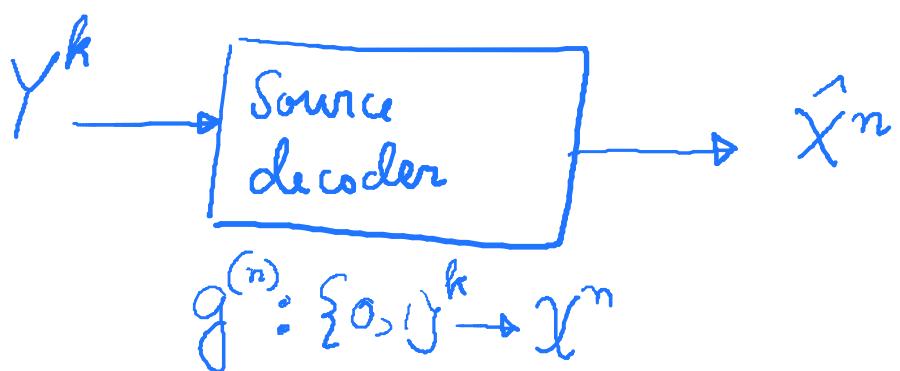
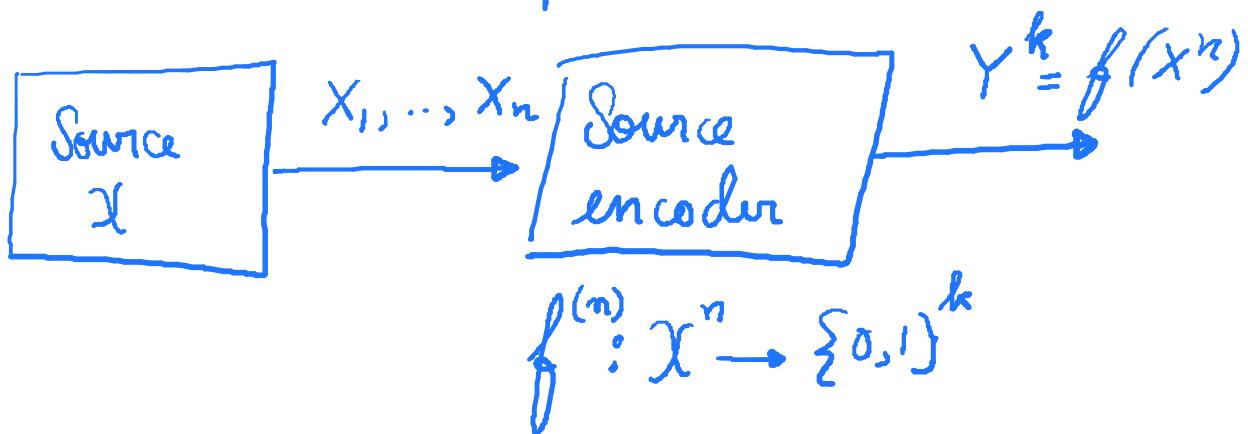
e.g. $\mathcal{X} = \{A, B, +, \rightarrow, Z, a, b, \dots, z, 1, 2, \dots, 0\}$

$\mathcal{X} = \{\text{Red, Blue, Green}\}$, etc.

How many bits are needed to store n symbols from \mathcal{X} ?

$$- n \lceil \log |\mathcal{X}| \rceil.$$

Towards formulating the lossless source-coding problem



Correct decoding constraint :-

$$\forall x^n \in X^n \quad g^{(n)}(f^{(n)}(x^n)) = x^n.$$

— This is why we call it
lossless.

We also call this setting fixed-length source coding since a fixed number of source symbols (n) gets mapped to a fixed number of bits.

Define rate $R = \frac{k}{n}$ of $f^{(n)} - g^{(n)}$ pair.

Easy to see :-

Any encoder-decoder pair

satisfying the correct decoding constraint

Must have a rate of at least $\log |X|$.

Why?

- To infer $|X|^n$ symbols, we need at least $n \log |X|$ bits.

Thus simply representing each of the symbols by their $\log |X|$ bit equivalents and then concatenating them is optimal in terms of rate.

- We do not have any compression here.

Road to compression - Redundancy

Suppose $S \subseteq X^n$ s.t. we do

not care about reconstructing

sequences for S .

Correct-Decoding Constraint :-

$$x^n = g^{(n)}(f^{(n)}(x^n)), \quad \forall x^n \in S^c.$$

Now, we need to ensure only sequences in S^c can be reconstructed.

This needs $\log |S|^c = \log (|X|^n - |S|)$ bits.

Thus, rate must be at least

$$\frac{1}{n} \log (|X|^n - |S|).$$

This rate is indeed achievable by using this simple scheme.

Using $f^{(n)}$ map every sequence in S^c to

strings of $\log(|X|^n - |S|)$ bits in 1-1 fashion. Let f^n map sequences in S to the all zeros string.

$$g^{(n)} : \{0,1\}^{\log(|X|^n - |S|)} \rightarrow X^n \text{ is chosen}$$

to map back the strings of $\log(|X|^n - |S|)$ bits to their preimages under f^n in S^c .

It is easy to see that

$$\frac{1}{n} \log(|X|^n - |S|) < \log |X|,$$

and hence we have compression.

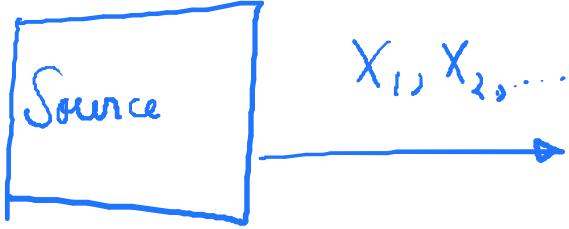
Note :- i) In reality S may not be known beforehand.

ii) It may be that $S = \emptyset$, but some elements of $|X|^m$ rarely occur.

Solution :- Use probability to model the source

The actual setup

- Lossless fixed-length source coding for discrete memoryless sources



Alphabet X_i

Pmf P_X

X_1, X_2, \dots

$X_i \sim P_X$ i.i.d.

Probabilistic correct decoding :-

$$P_X \left(g^{(n)}(f^{(n)}(X^n)) \neq X^n \right) \leq \varepsilon$$

for some fixed $\varepsilon > 0$.

Theorem (Löwles source coding theorem) :-

Consider a discrete memoryless source

with alphabet X and pmf P_X , and

fix $\varepsilon > 0$. There exists a

sequence of encoder-decoder pairs

$(f^{(n)}, g^{(n)})_{n \geq 1}$, $f^{(n)}: \mathcal{X}^n \rightarrow \{0,1\}^k$,

$g^{(n)}: \{0,1\}^k \rightarrow \mathcal{N}^n$, satisfying the

probabilistic correct decoding constraint

such that the rate $R^{(n)} = \frac{k}{n}$

satisfies $R^{(n)} \leq H(P_x) + \varepsilon$.

Furthermore, any encoder-decoder pair sequence of rate $R'^{(n)}$ satisfying the probabilistic correct decoding

constraint must satisfy $R'^{(n)} \geq H(P_x) - \varepsilon$.

Here $H(P_x) \triangleq \sum_{x \in \mathcal{X}} P_x(x) \log \frac{1}{P_x(x)}$,

is called the entropy of

P_x or the entropy of the source

The notation $a_n \leq_n b_n$ means $\exists N$ s.t.

$a_n \leq b_n \forall n \geq N$. Similarly the notation $a_n \geq_n b_n$ is defined.

Notes :-

- i) In layman's terms the above theorem is summarized as "The least rate of source coding possible is $H(P_x)$ and there exist a source coding scheme achieving it".
- ii) Notice that we do not call entropy 'a measure of randomness'. Instead, what we

show is that entropy is a quantity which equals to an operational quantity, 'the least rate of compression'. This operational quantity, i.e., the least rate of compression can then be interpreted as the amount of uncertainty/information in the source.

and hence you cannot compress beyond this point.

* Proof of the lossless source coding theorem

→ The asymptotic equipartition property

Fix any $n \geq 1$ and $\varepsilon > 0$.

Define the set $A_\varepsilon^{(n)} \subseteq \mathcal{X}^n$ as

$$A_\varepsilon^{(n)} := \left\{ x^n \in \mathcal{X}^n : \left| \frac{1}{n} \sum_{i=1}^n \log \frac{1}{P_x(x_i)} - H(P_x) \right| \leq \varepsilon \right\}$$

The set $A_\varepsilon^{(n)}$ is called the set of typical sequences.

• Lemma :- If $X \sim P_X$, then

$$\mathbb{E} \left[\log \frac{1}{P_x(x)} \right] = H(X).$$

$$\text{Proof: } \mathbb{E} \left[\log \frac{1}{P_x(x)} \right]$$

$$= \sum_{x \in X} P_X(x) \log \frac{1}{P_X(x)} = H(P_X) \quad \dots \quad \blacksquare$$

• Theorem (Asymptotic Equipartition Property) :- The typical set $A_\varepsilon^{(n)}$

satisfies

$$i) P(A_\varepsilon^{(n)}) \rightarrow 1.$$

$$ii) |A_\varepsilon^{(n)}| \leq 2^{n(H(P_x) + \varepsilon)}$$

$$iii) |A_\varepsilon^{(n)}| \geq (1-\varepsilon) \cdot 2^{n(H(P_x) - \varepsilon)}$$

Proof :- i) Define $Y_i = \log \frac{1}{P_x(x_i)}$

Then Y_i are i.i.d.

Thus, by the WLLN,

$$\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\text{P}} \mathbb{E}[Y_i] = H(P_x).$$

$$\Rightarrow P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \frac{1}{P_X(x_i)} - H(P_X)\right| > \varepsilon\right) \xrightarrow{\quad} 0$$

Thus, $P\left(\left|\frac{1}{n} \sum_{i=1}^n \log \frac{1}{P_X(x_i)} - H(P_X)\right| \leq \varepsilon\right) \xrightarrow{\quad} 1,$
 i.e., $P(A_\varepsilon^{(n)}) \rightarrow 1.$

ii) Let $x^n \in A_\varepsilon^{(n)}$.

$$n(H(P_X) - \varepsilon) \leq \log \left(\frac{1}{\prod_{i=1}^n P_X(x_i)} \right) \leq n(H(P_X) + \varepsilon)$$

$$\Rightarrow n(H(P_X) - \varepsilon) \leq \log \frac{1}{P_X(x^n)} \leq n(H(P_X) + \varepsilon)$$

$$\Rightarrow 2^{-n(H(P_X) + \varepsilon)} \leq P_X(x^n) \leq 2^{-n(H(P_X) - \varepsilon)}$$

. . . . (A)

Now, $P(A_{\varepsilon}^{(n)}) \leq 1$

$$\Rightarrow \sum_{x^n \in A_{\varepsilon}^{(n)}} P_x(x^n) \leq 1$$

$$\Rightarrow |A_{\varepsilon}^{(n)}| \cdot 2^{-n(H(P_x) + \varepsilon)} \leq 1 \quad [\text{From } \textcircled{A}]$$

$$\Rightarrow |A_{\varepsilon}^{(n)}| \leq 2^{n(H(P_x) + \varepsilon)}$$

iii) $\therefore P(A_{\varepsilon}^{(n)}) \rightarrow 1$

$$P(A_{\varepsilon}^{(n)}) \geq_n (1 - \varepsilon)$$

Thus, $\sum_{x^n \in A_{\varepsilon}^{(n)}} P_x(x^n) \geq_n (1 - \varepsilon)$

$$\Rightarrow |A_{\varepsilon}^{(n)}| 2^{-n(H(P_x) - \varepsilon)} \geq_n (1 - \varepsilon)$$

[\text{From } \textcircled{A}]

$$\Rightarrow |A_{\varepsilon}^{(n)}| \geq_n (1 - \varepsilon) \cdot 2^{n(H(P_x) - \varepsilon)}$$

... \textcircled{B}

→ Proof of the achievability part
of the source coding theorem

To show:- There exists an $(f^{(n)} \circ g^{(n)})_{n \geq 1}$
satisfying $R^{(n)} \leq H(P_x) + \epsilon.$

Define $f^{(n)}: X^n \rightarrow \{0, 1\}^k$ as follows.

Map each element of $A_\epsilon^{(n)}$ to a
different element of $\{0, 1\}^k$.

Map elements in $X^n \setminus A_\epsilon^{(n)}$ to 0^k.

We need $k = \log |A_\epsilon^{(n)}| \leq n(H(P_x) + \epsilon)$
[By AEP]

Define $g^{(n)}: \{0, 1\}^k \rightarrow X^n$ as follows.

Every element in $\{0,1\}^k$ is mapped to its corresponding element in $A_\varepsilon^{(n)}$ (according to $f^{(n)}$).

$$\begin{aligned} \text{Thus, } P_x(g^{(n)}(f^{(n)}(x^n)) \neq x^n) \\ &\leq P_x(A_\varepsilon^{(n)} c) \\ &\leq_n \varepsilon \quad [\because P_x(A_\varepsilon^{(n)}) \rightarrow 1] \end{aligned}$$

Thus, for all sufficiently large n , $f^{(n)}, g^{(n)}$ is a valid encoder-decoder pair.

Also, $R^{(n)} = \frac{k}{n} \leq H(P_x) + \varepsilon$ as required.

→ Proof of the converse part of the lossless source coding theorem

To show that any valid sequence of encoder-decoder pairs $(f^{(n)}, g^{(n)})$ must satisfy $R^{(n)} \geq_n H(P_x) - \varepsilon$.

Let $(f^{(n)}, g^{(n)})$ be a valid sequence of encoder-decoder pairs.

Let $B^{(n)} = \{x^n : g^{(n)}(f^{(n)}(x^n)) = x^n\}$

Then, the validity of $(f^{(n)}, g^{(n)})$

implies that $P_x(B^{(n)}) \geq_n 1 - \varepsilon$

..... (A)

Now, from the definition of $B^{(n)}$, we must have $k \geq \log |B^{(n)}|$

We shall now bound $P(B^{(n)})$.

To do so, observe that

$$\begin{aligned} & P(A_{\frac{\varepsilon}{2}}^{(n)} \cap B^{(n)}) \\ &= 1 - P(A_{\frac{\varepsilon}{2}}^{(n)c} \cup B^{(n)c}) \\ &\geq 1 - P(A_{\frac{\varepsilon}{2}}^{(n)c}) - P(B^{(n)c}) \\ &\stackrel{n}{\geq} 1 - \frac{3}{2}\varepsilon \quad [\text{Union Bound}] \\ &\quad [\because P(A_{\frac{\varepsilon}{2}}^{(n)}) \rightarrow 1, \\ &\quad P(B^{(n)}) \stackrel{n}{\geq} 1 - \varepsilon] \\ &\quad \dots \quad \textcircled{B} \end{aligned}$$

On the other hand,

$$\begin{aligned} P(A_{\frac{\varepsilon}{2}}^{(n)} \cap B^{(n)}) &= \sum_{x^n \in A_{\frac{\varepsilon}{2}}^{(n)} \cap B^{(n)}} P_X(x^n) \\ &\leq \sum_{x^n \in A_{\frac{\varepsilon}{2}}^{(n)} \cap B^{(n)}} 2^{-n(H(P_X) - \frac{\varepsilon}{2})} \quad (\because x^n \in A_{\frac{\varepsilon}{2}}^{(n)}) \end{aligned}$$

$$= |A_{\frac{\varepsilon}{2}}^{(n)} \cap B^{(n)}| \cdot 2^{-n(H(P_x) - \frac{\varepsilon}{2})}$$

$$\leq |B^{(n)}| \cdot 2^{-n(H(P_x) + \frac{\varepsilon}{2})} \quad \dots \quad \textcircled{B}$$

Combining \textcircled{A} and \textcircled{B} , we have

$$|B^{(n)}| \geq_n 2^{n(H(P_x) - \frac{\varepsilon}{2})} \cdot \left(1 - \frac{3}{2}\varepsilon\right)$$

$$\text{Thus, } h \geq \log |B^{(n)}|$$

$$\geq_n n \cdot \left(H(P_x) - \frac{\varepsilon}{2}\right) + \log \left(1 - \frac{3}{2}\varepsilon\right)$$

$$\text{Then, } R^{(n)} \geq_n H(P_x) - \frac{\varepsilon}{2} - \frac{1}{n} \log \frac{1}{\left(1 - \frac{3}{2}\varepsilon\right)}$$

$$\geq_n H(P_x) - \frac{\varepsilon}{2} - \frac{\varepsilon}{2}$$

$$= H(P_x) - \varepsilon.$$

$\dots \textcircled{C}$

* Entropy, conditional entropy, and its properties

Let $X \sim P_x$.

We shall interchangeably write

$H(X)$ and $H(P_x)$.

Joint entropy :-

Let $P_{x,y}$ be the joint pmf of X and Y .

Then

$$H(X, Y) \triangleq \sum_{x,y} P_{xy}(x,y) \cdot \log \frac{1}{P_{xy}(x,y)}$$

Conditional entropy :-

Let $X, Y \sim P_{XY} = P_{X|Y} \cdot P_Y$

For any y ,

$$H(X|Y=y) \triangleq \sum_x P_{X|Y}(x|y) \cdot \log \frac{1}{P_{X|Y}(x|y)}$$

The conditional entropy of X given Y is defined as

$$\begin{aligned} H(X|Y) &= \mathbb{E}_{P_Y} [H(X|Y=y)] \\ &= \sum_y P_Y(y) \cdot H(X|Y=y) \\ &= \sum_{x,y} P_{X,Y}(x,y) \cdot \log \frac{1}{P_{X|Y}(x|y)} \end{aligned}$$

→ Properties of entropy

i> $H(X) \geq 0$ with equality iff
 X is a constant

To see this, note that

$$H(X) = \sum_x P_x(x) \log \frac{1}{P_x(x)}$$

$$\log \frac{1}{P_x(x)} \geq 0, \quad P_x(x) \geq 0.$$

Thus $H(X) \geq 0$.

Also, $H(X) = 0$ iff

$$P_x(x) = 0 \text{ or } P_x(x) = 1, \quad \forall x$$

i.e., X is a constant.

ii)

Chain rule:

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i; X^{i-1})$$

To see this,

$$H(X_1, \dots, X_n) = \sum_{x^n} P_{X^n}(x^n) \log \frac{1}{P_{X^n}(x^n)}$$

$$= \sum_{x^n} P_{X^n}(x^n) \cdot \log \frac{1}{\prod_{i=1}^n P_{X_i | X^{i-1}}(x_i | x^{i-1})}$$

$$= \sum_{x^n} \sum_{i=1}^n P_{X^n}(x^n) \cdot \log \frac{1}{P_{X_i | X^{i-1}}(x_i | x^{i-1})}$$

$$= \sum_{i=1}^n \sum_{x^i} \sum_{\substack{x_{i+1}, \dots, \\ x_n}} P_{X^{i-1}}(x^{i-1}) \cdot P_{X_i | X^{i-1}}(x_i | x^{i-1}).$$

$$P_{X_{i+1}, \dots, X_n | X^i}(x_{i+1}, \dots, x_n | x^i).$$

$$\log \frac{1}{P_{X_i | X^{i-1}}(x_i | x^{i-1})}$$

$$= \sum_{i=1}^n \sum_{x^i} P_{X^{i-1}}(x^{i-1}) \cdot P_{X_i | X^{i-1}}(x_i | x^{i-1}).$$

$$\log \frac{1}{P_{X_i | X^{i-1}}(x_i | x^{i-1})}$$

$$= \sum_{i=1}^n H(X_i | X^{i-1}).$$

iii) Let f be any function of X .

Then, a) $H(f(X)|X)=0$

b) $H(f(X)) \leq H(X)$,

with equality iff
 f is 1-1.

To see this, let $Y=f(X)$.

Now, fix any x .

Then $P_{Y|X}(y|x) = 1 \{f(x)=y\}$

Thus, $H(Y|X=x)=0$.

Hence, $H(f(X)|X)=H(Y|X)$

$$= 0.$$

Now, expand $H(f(X), X)$ using
the chain rule.

$$H(f(x), x) = H(x) + H(f(x)|x) = H(x)$$

$$H(f(x), x) = H(f(x)) + H(x|f(x)).$$

Thus, $H(f(x)) = H(x) - H(x|f(x))$

Since $H(x|f(x)) \geq 0$,

we have

$$H(f(x)) \leq H(x).$$

Finally, we have

$$H(f(x)) = H(x)$$

$$\iff H(x|f(x)) = 0$$

$$\iff H(x|f(x)=y) = 0, \forall y \in \text{range}(f)$$

$$\iff x \text{ is a constant if } f(x)=y, \forall y \in \text{range}(f)$$

$$\iff f \text{ is 1-1.}$$

* Hypothesis testing

Consider two hypotheses

$$H_0 : X^n \sim \text{i.i.d } P_0$$

$$H_1 : X^n \sim \text{i.i.d } P_1$$

$$P(H_0) = 1 - P(H_1) = \rho_0$$

Goal :- Observe X^n and decide

H_0 or H_1 .

A hypothesis test is a function

$$g : \mathcal{X}^n \rightarrow \{0, 1\}.$$

$$\text{Define } \Gamma = \{x^n : g(x^n) = 1\}$$

How good is a hypothesis test?

- Probability of missed detection

$$P_M \triangleq P(g(X^n) = 0 | H_1) = P_1(\Gamma^c)$$

- Probability of false alarm

$$P_{FA} \triangleq P(g(X^n) = 1 | H_0) = P_0(\Gamma)$$

- Probability of error

$$P_e = p_0 P_{FA} + (1-p_0) \cdot P_M$$

* Stein's lemma and relative entropy /

KL divergence

Let P_0, P_1 be any two pmfs on X .

Define,

$$D(P_0 || P_1) \triangleq \sum_{x \in X} P_0(x) \log \frac{P_0(x)}{P_1(x)}$$

$D(\cdot||\cdot)$ is known as the KL divergence or relative entropy.

Note :- $D(\cdot||\cdot)$ is not symmetric.

i.e., $D(P_0||P_1)$ is not necessarily same as $D(P_1||P_0)$. Still it can be viewed as a 'distance' between probability measures, as we shall see below.

• Lemma (Stein's lemma) :-

Consider a pair of hypothesis

$$H_0: X^n \sim \text{i.i.d. } P_0$$

$$H_1: X^n \sim \text{i.i.d. } P_1$$

with prior $p_0 = P(H_0) = 1 - P(H_1)$.

Fix any $0 < \alpha < \frac{1}{2}$.

Let $\tilde{g}^{(n)}: \mathcal{X}^n \rightarrow \{0, 1\}$ be a sequence of hypothesis tests with decision regions $\Gamma^{*(n)}$ satisfying the following:

$$P_1(\Gamma^{*(n)c}) = \min_{\Gamma^{(n)} \subseteq \mathcal{X}^n:} P_1(\Gamma^{(n)c})$$

$$P_0(\Gamma^{(n)}) < \alpha.$$

Then, $\lim_{n \rightarrow \infty} \frac{1}{n} \log P_1(\Gamma^{*(n)c}) = -D(P_0 \parallel P_1).$

- Note :- 1. Since we are dealing with discrete entities, we are good to use 'min' instead of 'inf':
2. The lemma is saying the following. Among all possible hypothesis tests with $P_{FA} < \epsilon$,

$P_M \rightarrow 0$ as $n \rightarrow \infty$ with an exponent
 $-D(P_0 || P_1)$, i.e. $P_M \approx 2^{-n \cdot D(P_0 || P_1)}$.

3. Heuristic :- P_M will go down to zero
quicker if P_0 is 'further away'
from P_1 . Thus, the exponent
 $D(P_0 || P_1)$ is in some sense
quantifying the 'distance'
between P_0 and P_1 .

Lemma (AEP for relative entropy) :-

Define $A_\varepsilon^{(n)}(P_0 || P_1) = \left\{ x^n : \left| \frac{1}{n} \log \frac{P_0(x^n)}{P_1(x^n)} - D(P_0 || P_1) \right| \leq \varepsilon \right\}$,

for any $\varepsilon > 0$. Then,

$$(i) P_0(A_\varepsilon^{(n)}(P_0 || P_1)) \rightarrow 0$$

$$(ii) P_1(A_\varepsilon^{(n)}(P_0 || P_1)) \leq 2^{-n(D(P_0 || P_1) - \varepsilon)}$$

$$(iii) P_1(A_\varepsilon^{(n)}(P_0 || P_1)) \geq (1 - \varepsilon) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)}$$

Proof:- (i) Define $Y_i = \log \frac{P_0(X_i)}{P_1(X_i)}$

$$\text{Then, } \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \log \frac{P_0(x^n)}{P_1(x^n)}.$$

Now, noting that given H_0 ,

Y_1, Y_2, \dots are i.i.d., we have

by the WLLN

$$P_0 \left(\left| \frac{1}{n} \sum_{i=1}^n Y_i - E_0[Y_i] \right| > \varepsilon \right) \rightarrow 0$$

$$\Rightarrow P_0 \left(\left| \frac{1}{n} \log \frac{P_0(x^n)}{P_1(x^n)} - D(P_0 || P_1) \right| > \varepsilon \right) \rightarrow 0$$

$$\Rightarrow P_0(A_\varepsilon^{(n)}(P_0 || P_1)) \rightarrow 0.$$

(ii) If $x^n \in A_\varepsilon^{(n)}(P_0 || P_1)$,
then

$$\left| \frac{1}{n} \log \frac{P_0(x^n)}{P_1(x^n)} - D(P_0 || P_1) \right| \leq \varepsilon$$

$$\Rightarrow 2^{n(D(P_0 || P_1) - \varepsilon)} \leq \frac{P_0(x^n)}{P_1(x^n)} \leq 2^{n(D(P_0 || P_1) + \varepsilon)}$$

$$\Rightarrow P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)} \leq P_1(x^n) \leq P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) - \varepsilon)}$$

$$\text{Now, } P_1(A_\varepsilon^{(n)}(P_0 || P_1))$$

$$= \sum_{x^n \in A_\varepsilon^{(n)}(P_0 || P_1)} P_1(x^n)$$

$$\leq \sum_{x^n \in A_\varepsilon^{(n)}(P_0 || P_1)} P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) - \varepsilon)}$$

$$= 2^{-n(D(P_0 || P_1) - \varepsilon)} \cdot P_0(A_\varepsilon^{(n)}(P_0 || P_1))$$

$$\leq 2^{-n(D(P_0 || P_1) - \varepsilon)}$$

$$\text{iii) } P_1(A_\varepsilon^{(n)}(P_0 || P_1)) = \sum_{x^n \in A_\varepsilon^{(n)}(P_0 || P_1)} P_1(x^n)$$

$$\geq \sum_{x^n \in A_\varepsilon^{(n)}(P_0 || P_1)} P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)}$$

$$= 2^{-n(D(P_0 || P_1) + \varepsilon)} \cdot P_0(A_\varepsilon^{(n)}(P_0 || P_1))$$

$$\geq_n (1 - \varepsilon) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)}$$

[By part (i)]

... · \emptyset

- Proof of Stein's lemma

Fix any $\varepsilon > 0$.

We will show that there exist

a test with decision region $\Gamma^{(n)}$

s.t. $P_0(\Gamma^{(n)}) \leq_n \alpha$ and

$\frac{1}{n} \log P_1(\Gamma^{(n)}) \leq -D(P_0 || P_1) + \varepsilon$

Then, noting that \log is an increasing function, we will have

$$\frac{1}{n} \log P_i(\Gamma^{*(n)}) \leq -D(P_0 || P_i) + \varepsilon. \quad \dots \quad (a)$$

We see that the above is true by simply choosing $\Gamma^{(n)} = \hat{\Gamma}_\varepsilon^{(n)}(P_0 || P_i)$, and then noting that $P_0(\Gamma^{(n)}) \leq \alpha$ and

$$P_i(\Gamma^{*(n)}) \leq 2^{-n(D(P_0 || P_i) - \varepsilon)} \text{ from the AEP.}$$

Next, let $\Gamma'^{(n)}$ be any decision region satisfying $P_0(\Gamma'^{(n)}) \leq \alpha$.

Then, $P_i(\Gamma'^{(n)})$

$$= \sum_{x^n \in \Gamma'^{(n)}} P_i(x^n)$$

$$\geq \sum_{\substack{x^n \in A_\varepsilon^{(n)}(P_0 || P_1) \\ \cap \Pi'^{(n)}}} P_1(x^n)$$

$$\geq \sum_{\substack{x^n \in A_\varepsilon^{(n)}(P_0 || P_1) \cap \Pi'^{(n)}}} P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)}$$

$$[\because P_1(x^n) \geq P_0(x^n) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)}$$

$$= P_0(A_\varepsilon^{(n)}(P_0 || P_1) \cap \Pi'^{(n)}) \cdot 2^{-n(D(P_0 || P_1) + \varepsilon)} \quad \text{if } x^n \in A_\varepsilon^{(n)}(P_0 || P_1)$$

Next, note that $P_0(A_\varepsilon^{(n)}(P_0 || P_1))^c \leq_n \alpha$
 (By (i) of AEP)

$$\text{Then, } P_0(A_\varepsilon^{(n)}(P_0 || P_1) \cap \Pi'^{(n)})$$

$$= 1 - P_0(A_\varepsilon^{(n)}(P_0 || P_1)^c \cup \Pi'^{(n)})$$

$$\geq 1 - P_0(A_\varepsilon^{(n)}(P_0 || P_1)^c) - P_0(\Pi'^{(n)})$$

$$\geq_n 1 - 2\alpha$$

$$\text{Thus, } P_1(\pi^{(n)}_{1^c}) \geq_n (1-\alpha) \cdot 2^{-n(D(P_0||P_1) + \varepsilon)}$$

$$\begin{aligned} \text{Thus, } \frac{1}{n} \log P_1(\pi^{(n)}_{1^c}) &\geq_n \frac{1}{n} \log (1-\alpha) \\ &\quad - D(P_0||P_1) - \varepsilon \\ &\geq_n - D(P_0||P_1) - 2\varepsilon \end{aligned}$$

Since $\pi^{*(n)}$ satisfies $P_0(\pi^{*(n)}) \leq_n \alpha$,

we have

$$\begin{aligned} \frac{1}{n} \log P_1(\pi^{*(n)}_{1^c}) &\geq_n - D(P_0||P_1) - 2\varepsilon \\ &\dots \textcircled{b} \end{aligned}$$

Combining \textcircled{a} and \textcircled{b}, we have

$$\left| \frac{1}{n} \log P_1(\pi^{*(n)}_{1^c}) + D(P_0||P_1) \right| \leq_n 2\varepsilon$$

Noting that the choice of ε was arbitrary, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P_i(P_i^{(n)}) = -D(P_0 || P_i)$$

..... \square

* Introduction to convex functions

• Definition:- Let S be a set with operations + under which it is closed and an operation with reals, \cdot , such that $c \cdot s \in S$, $\forall c \in \mathbb{R}, s \in S$.

A subset $X \subseteq S$ is called convex if for every $x_1, x_2 \in X$, and every $\lambda \in [0, 1]$,

$$\lambda \cdot x_1 + (1-\lambda) \cdot x_2 \in X.$$

e.g. Closed and open intervals are convex. Set of all pmf on finite set \mathcal{X} is convex.

• Definition :- Let S be a convex set.

A function $f: S \rightarrow \mathbb{R}$ is called convex

if for every $s_1, s_2 \in S$ and every $\lambda \in [0, 1]$

$$\lambda f(s_1) + (1-\lambda) f(s_2) \geq f(\lambda s_1 + (1-\lambda) s_2).$$

• Lemma :- A function $f: [a, b] \rightarrow \mathbb{R}$

which is differentiable is convex

iff $f'(x)$ is increasing.

Proof :- First assume that

f is convex. We shall first prove that for any

$$x_1 < x < x_2 \in [a, b],$$

we have $\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}$

$$\leq \frac{f(x_2) - f(x)}{x_2 - x} .$$

..... @

If @ is true then let $h \in [0, x_2 - x_1]$,

and observe that

$$\frac{f(x_1 + h) - f(x_1)}{h} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

and hence taking limit $h \rightarrow 0^+$

$$f'(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

Similarly, we have using @

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_2 - h) - f(x_2)}{-h},$$

which upon taking limit $h \rightarrow 0^+$

give $\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f'(x_2)$.

Combining we get $f'(x_1) \leq f'(x_2)$ as needed.
 Thus to complete the only if part,
 we need to prove @

To do so, define $\lambda = \frac{x - x_1}{x_2 - x_1}$

$$\begin{aligned} \text{Now, } f(x) &= f(x_1 + \lambda \cdot (x_2 - x_1)) \\ &= f(\lambda x_2 + (1-\lambda) \cdot x_1) \\ &\leq \lambda f(x_2) + (1-\lambda) \cdot f(x_1) \end{aligned}$$

Thus,

$$\begin{aligned} (x_2 - x_1) f'(x) &\leq (x - x_1) f(x_2) \\ &\quad + (x_2 - x) \cdot f(x_1) \end{aligned}$$

$$\Rightarrow (x_2 - x_1) [f(x) - f(x_1)] \leq (x - x_1) (f(x_2) - f(x_1))$$

$$\Rightarrow \frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Also,

$$(x_2 - x_1)f(x) \leq (x - x_1) \cdot f(x_2) + (x_2 - x)f(x)$$

$$\Rightarrow (x_2 - x_1)(f(x) - f(x_2)) \leq (x_2 - x)(f(x_1) - f(x_2))$$

$$\Rightarrow (x_2 - x_1)(f(x_2) - f(x)) \geq (x_2 - x)(f(x_2) - f(x_1))$$

$$\Rightarrow \frac{f(x_2) - f(x)}{x_2 - x} \geq \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

To prove the if part, assume $f'(x)$ is increasing.

Fix $x_1, x_2 \in [a, b]$ and $\lambda \in [0, 1]$, and

let $x = \lambda x_1 + (1-\lambda)x_2$.

By the mean value theorem, there exist

$y_1 \in (x_1, x)$, $y_2 \in (x, x_2)$, s.t.

$$f'(y_1) = \frac{f(x) - f(x_1)}{x - x_1}$$

$$f'(y_2) = \frac{f(x_2) - f(x)}{x_2 - x}$$

Now, since f' is increasing,

$$f'(y_2) \geq f'(y_1)$$

$$\Rightarrow \frac{f(x_2) - f(x)}{x_2 - x} \geq \frac{f(x) - f(x_1)}{x - x_1}$$

$$\begin{aligned} \Rightarrow (x - x_1)f(x_2) + (x_2 - x)f(x_1) \\ \geq f(x)(x_2 - x_1) \end{aligned}$$

$$\begin{aligned} \Rightarrow \left(\frac{x - x_1}{x_2 - x_1} \right) \cdot f(x_2) + \left(\frac{x_2 - x}{x_2 - x_1} \right) \cdot f(x_1) \\ \geq f(x) \end{aligned}$$

Plugging in $x = \lambda x_1 + (1-\lambda)x_2$, we get

$$(1-\lambda)f(x_0) + \lambda f(x_1) \geq f(\lambda x_0 + (1-\lambda)x_1).$$

Noting that the choice of x_0, x_1, λ , were arbitrary, we have that f is convex.

. . . 

* Jensen's Inequality

• Lemma (Jensen's inequality) :-

let $f: S \rightarrow \mathbb{R}$ be a convex function

and let X be an S -valued

random variable. Then,

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

For strictly convex f , the inequality holds with equality iff X is constant.

Proof when X takes only finite values in S :-

We prove by induction on the $|\text{range}(X)|$.

Let $|\text{range}(X)| = 2$, i.e., X takes

values in $\{x_1, x_2\} \subseteq S$.

Then, $E[f(X)] = \lambda f(x_1) + (1-\lambda) f(x_2)$
where $\lambda \triangleq P(x_1)$.

Then, by convexity of f

$$\begin{aligned} E[f(X)] &\geq f(\lambda x_1 + (1-\lambda)x_2) \\ &= f(E[X]). \end{aligned}$$

Next, assume that the inequality is true when $|\text{range}(X)| \leq n$.

Next, consider $|\text{range}(x)| = n+1$, i.e.

$$\text{range}(x) = \{x_1, x_2, \dots, x_{n+1}\}$$

where $P(x_i) = \lambda_i$.

We need to show

$$\sum_{i=1}^{n+1} \lambda_i f(x_i) \geq f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right).$$

$$\sum_{i=1}^{n+1} \lambda_i f(x_i)$$

$$= \sum_{i=1}^{n-1} \lambda_i f(x_i) + \lambda_n f(x_n) + \lambda_{n+1} f(x_{n+1})$$

$$= \sum_{i=1}^{n-1} \lambda_i f(x_i) + (\lambda_n + \lambda_{n+1}) \cdot \left[\frac{\lambda_n}{\lambda_n + \lambda_{n+1}} f(x_n) \right.$$

$$\left. + \frac{\lambda_{n+1}}{\lambda_n + \lambda_{n+1}} f(x_{n+1}) \right]$$

$$\geq \sum_{i=1}^{n-1} \lambda_i f(x_i) + (\lambda_n + \lambda_{n+1}) \cdot f\left(\frac{\lambda_n x_n}{\lambda_n + \lambda_{n+1}}$$

$$\left. + \frac{\lambda_{n+1} \cdot x_{n+1}}{\lambda_n + \lambda_{n+1}}\right)$$

[By convexity of f]

$$= E[f(Y)],$$

where $Y = X_i$, if $X = x_1, x_2, \dots, x_{n-1}$

$$= \frac{\lambda_n x_n}{\lambda_n + \lambda_{n+1}} + \frac{\lambda_{n+1} x_{n+1}}{\lambda_n + \lambda_{n+1}}, \text{ if } X = x_n, x_{n+1}$$

$$X = x_n, x_{n+1}$$

and $P(Y = x_i) = \lambda_i, \forall i \leq n-1$

$$P\left(Y = \frac{\lambda_n x_n}{\lambda_n + \lambda_{n+1}} + \frac{\lambda_{n+1} x_{n+1}}{\lambda_n + \lambda_{n+1}}\right)$$
$$= \lambda_n + \lambda_{n+1}$$

$|\text{range}(Y)| = n$. Then, by induction

hypothesis, we have

$$E[f(Y)] \geq f(E[Y])$$

$$= f\left(\sum_{i=1}^{n-1} \lambda_i x_i + (\lambda_n + \lambda_{n+1}) \cdot \left[\frac{x_n \cdot \lambda_n}{\lambda_n + \lambda_{n+1}} + \frac{x_{n+1} \cdot \lambda_{n+1}}{\lambda_n + \lambda_{n+1}} \right]\right)$$

$$= f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right)$$

Hence, $\sum_{i=1}^n \lambda_i f(x_i) \geq f\left(\sum_{i=1}^n \lambda_i x_i\right)$.

Finally, noting f is strictly convex, all inequalities are equalities iff $x_1 = x_2 = \dots = x_n$.

* Log-sum inequality

Lemma (log-sum inequality) :-

Let X be a finite set and let

$f, g : X \rightarrow \mathbb{R}_{\geq 0}$. Then

$$\sum_{x \in X} f(x) \log \frac{f(x)}{g(x)} \geq \left(\sum_{x \in X} f(x) \right) \cdot \log \frac{\left(\sum_{x \in X} f(x) \right)}{\left(\sum_{x \in X} g(x) \right)}$$

with equality iff $\frac{f(x)}{g(x)} = c$.

Proof:- Firstly, consider the function

$$h(x) = x \log x.$$

$$h'(x) = \log x + \log e$$

$$h''(x) = \frac{1}{x} \log e$$

Thus, $h(x)$ is convex when $x \geq 0$.

Define the random variable

$$Y = \frac{f(x)}{g(x)}, \text{ and}$$

consider the pmf $P(X=x) = \frac{g(x)}{\sum_{x'} g(x')}$.

Note that the pmf is well-defined
as $g(x) \geq 0$.

Also, $Y \geq 0$, since f, g are
non-negative valued.

Then,

$$\sum_{x \in X} f(x) \cdot \log \frac{f(x)}{g(x)}$$

$$= \sum_{x \in X} g(x) \cdot \frac{f(x)}{g(x)} \cdot \log \frac{f(x)}{g(x)}$$

$$= \left(\sum_{x \in X} g(x) \right) \cdot E[Y \log Y]$$

$$\geq \sum_{x \in X} g(x) \cdot (E Y) \log (E Y)$$

[Jensen's inequality

and noting $Y \geq 0$]

$$= \left(\sum_{x \in X} g(x) \right) \cdot \left(\sum_{x \in X} \frac{g(x)}{\left(\sum_{x' \in X} g(x') \right)} \cdot \frac{f(x)}{g(x)} \right).$$

$$\log \left(\sum_{x \in X} \frac{g(x)}{\left(\sum_{x' \in X} g(x') \right)} \cdot \frac{f(x)}{g(x)} \right)$$

$$= \left(\sum_{x \in X} f(x) \right) \cdot \log \frac{\left(\sum_{x \in X} f(x) \right)}{\left(\sum_{x \in X} g(x) \right)}$$

Finally, note that we have

equality iff Jensen's inequality
is satisfied with equality, which
happens iff $\frac{f(x)}{g(x)} = c$.



* Properties of Entropy and Relative Entropy

• **Hemmi :-** $D(P_0 || P_1) \geq 0$ and $D(P_0 || P_1) = 0$

iff $P_0 = P_1$.

Proof :-

$$D(P_0 || P_1)$$

$$= \sum_{x \in X} P_0(x) \log \frac{P_0(x)}{P_1(x)}$$

$$\geq \left(\sum_{x \in X} P_0(x) \right) \log \frac{\left(\sum_{x \in X} P_0(x) \right)}{\left(\sum_{x \in X} P_1(x) \right)} \quad [\text{Log-sum inequality}]$$

$$= 0$$

Note that we have inequality iff

$$\exists c \in \mathbb{R} \text{ s.t. } \frac{P_0(x)}{P_1(x)} = c, \quad \forall x \in X.$$

Noting that P_0, P_1 are pmfs

then $c = 1$, i.e., $P_0 = P_1$.

..... \square

- **Lemma :-** Let P_x be a pmf on \mathcal{X} . Then $H(P_x) \leq \log |\mathcal{X}|$ with equality iff P_x is uniform.

Proof:-

$$\begin{aligned}
 & \log |\mathcal{X}| - H(P_x) \\
 &= \sum_{x \in \mathcal{X}} P(x) \log |\mathcal{X}| - \sum_{x \in \mathcal{X}} P_x(x) \log \frac{1}{P_x(x)} \\
 &= \sum_{x \in \mathcal{X}} P_x(x) \cdot \log |\mathcal{X}| \cdot P_x(x) \\
 &= \sum_{x \in \mathcal{X}} P_x(x) \log \frac{P_x(x)}{1/\mathcal{X}} \\
 &= D(P_x \parallel \text{unif}\{\mathcal{X}\}) \\
 &\geq 0.
 \end{aligned}$$

Equality occurs iff $D(P_x \parallel \text{unif}\{\mathcal{X}\}) = 0$

$\Leftrightarrow P_x = \text{unif}\{x\}$.

B

• **Lemma (conditioning reduces entropy) :-**

$H(X) \geq H(X|Y)$, with equality iff
 $X \perp Y$.

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

$$= \sum_{x \in X} P_X(x) \cdot \log \frac{1}{P_X(x)}$$

$$- \sum_{x \in X} \sum_{y \in Y} P_{XY}(x, y) \cdot \log \frac{1}{P_{X|Y}(x|y)}$$

$$= \sum_{x \in X} \sum_{y \in Y} P_{XY}(x, y) \cdot \log \frac{1}{P_X(x)}$$

$$- \sum_{x \in X} \sum_{y \in Y} P_{XY}(x, y) \cdot \log \frac{1}{P_{X|Y}(x|y)}$$

$$= \sum_{y \in Y} \sum_{x \in X} P_{X,Y}(x, y) \log \frac{P_{X|Y}(x|y)}{P_X(x)}$$

$$= \sum_{y \in Y} P_Y(y) \cdot D(P_{X|Y}(\cdot|y) \| P_X | Y=y)$$

$$\geq 0$$

We have inequality iff $\forall y$

$$D(P_{X|Y}(\cdot|y) \| P_X | Y=y) = 0$$

$$\Leftrightarrow P_{X|Y}(\cdot|y) = P_X$$

$$\Leftrightarrow X \perp Y.$$

□

• Lemma :- $\sum_{i=1}^n H(X_i) \geq H(X^n)$,

with equality iff X_1, \dots, X_n are mutually independent

Proof:- By the chain rule we have

$$\begin{aligned}
 & H(X^n) \\
 &= \sum_{i=1}^n H(x_i | X^{i-1}) \\
 &\leq \sum_{i=1}^n H(x_i) \quad [\because \text{Conditioning reduces entropy}]
 \end{aligned}$$

Equality occurs iff $H(x_i | X^{i-1}) = H(x_i)$

$$\forall 1 \leq i \leq n$$

$$\iff X_i \perp X^{i-1} \quad \forall 1 \leq i \leq n$$

$\iff X_1, \dots, X_n$ are mutually independent

..... \square

• Conditional relative entropy

Consider two sets X and Y .

Let P_Y be a pmf on Y .

Let $P_{X|Y}^{(\cdot|y)}, Q_{X|Y}^{(\cdot|y)}$ be two conditional pmfs on X given $Y=y$.

Define

$$D(P_{X|Y}(\cdot|y) \parallel Q_{X|Y}(\cdot|y) | Y=y)$$

$$\triangleq \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \cdot \log \frac{P_{X|Y}(x|y)}{Q_{X|Y}(x|y)}$$

Define,

$$D(P_{X|Y} \parallel Q_{X|Y} | P_Y)$$

$$\triangleq \sum_{y \in \mathcal{Y}} P_Y(y) \cdot D(P_{X|Y}(\cdot|y) \parallel Q_{X|Y}(\cdot|y) | Y=y)$$

$$= \mathbb{E}_{P_Y} D(P_{X|Y}(\cdot|Y) \parallel Q_{X|Y}(\cdot|Y) | Y)$$

• Lemma (Convexity of relative entropy):-

Consider pmfs P_1, Q_1, P_2, Q_2 and some $\lambda \in [0, 1]$. Then,

$$\lambda D(P_1 \parallel Q_1) + (1-\lambda) D(P_2 \parallel Q_2)$$

$$\geq D(\lambda P_1 + (1-\lambda) P_2 \parallel \lambda Q_1 + (1-\lambda) Q_2).$$

Proof :- Firstly note that
 $\lambda P_1 + (1-\lambda) P_2$ and $\lambda Q_1 + (1-\lambda) Q_2$
are also pmf, and hence the statement of the lemma is well-defined.

Now,

$$\begin{aligned}
& \lambda D(P_1 || Q_1) + (1-\lambda) D(P_2 || Q_2) \\
&= \sum_{x \in X} \left(\lambda P_1(x) \log \frac{P_1(x)}{Q_1(x)} + (1-\lambda) \cdot P_2(x) \cdot \log \frac{P_2(x)}{Q_2(x)} \right) \\
&= \sum_{x \in X} \left(\lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda Q_1(x)} + (1-\lambda) P_2(x) \log \frac{(1-\lambda) \cdot P_2(x)}{(1-\lambda) \cdot Q_2(x)} \right) \\
&\geq \sum_{x \in X} (\lambda P_1(x) + (1-\lambda) P_2(x)) \cdot \log \frac{\lambda P_1(x) + (1-\lambda) P_2(x)}{\lambda Q_1(x) + (1-\lambda) Q_2(x)} \\
&\quad \left[\text{Log-sum inequality} \right] \\
&= D(\lambda P_1 + (1-\lambda) P_2 || \lambda Q_1 + (1-\lambda) Q_2).
\end{aligned}$$

• Lemma (concavity of entropy) :-

$H(P_x)$ is a concave function of P_x .

Proof:- $H(P_x) = \log |X| - D(P_x || \text{unif } \{x\})$

Now, choose P_{x_1}, P_{x_2} and any $\lambda \in [0, 1]$.

Note that $\lambda \text{unif } \{x\} + (1-\lambda) \text{unif } \{x\}$
 $= \text{unif } \{x\}$.

Then,

$$\lambda H(P_{x_1}) + (1-\lambda) \cdot H(P_{x_2})$$

$$= \log |X| (\lambda + (1-\lambda))$$

$$= \lambda D(P_{x_1} || \text{unif } \{x\}) + (1-\lambda) D(P_{x_2} || \text{unif } \{x\})$$

$$\leq \log |X| - D(\lambda P_{x_1} + (1-\lambda) P_{x_2} || \text{unif } \{x\})$$

$$= H(\lambda P_{x_1} + (1-\lambda) P_{x_2}).$$

Thus, entropy is concave

.. \square

• Binary entropy:-

Let $X \sim Be(p)$.

Then $h(p) \triangleq H(X)$

i.e.,

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

Note that $h(0), h(1)$ are technically undefined, but

$$\lim_{p \rightarrow 0^+} h(p) = \lim_{p \rightarrow 1^-} h(p) = 0.$$

Thus we define $h(0) = h(1) = 0$.

Lemma :-

i) $h(p) = h(1-p)$

ii) $h(p)$ is continuous for $p \in [0,1]$.

iii) $\max_{p \in [0,1]} h(p) = 1$ which occur at $p = \frac{1}{2}$.

Proof :- i) Obvious from definition.

ii) Continuity for $h \in (0,1)$

follow by continuity of
log and linear functions.

$$\text{Since } h(0) = \lim_{f \rightarrow 0^+} h(f)$$

$$h(1) = \lim_{f \rightarrow 1^-} h(f)$$

$h(f)$ is continuous for $f \in [0,1]$.

iii) Noting that $[0,1]$ is compact

and $h(\cdot)$ is continuous,

range ($h(\cdot)$) is compact and

hence closed. Thus, a maximum exists.

To find the maxima,

$$h'(p) = \log e + \log \frac{1}{p}$$

$$= -\log e - \log \frac{t}{(1-p)}$$

$$= \log \left(\frac{1-p}{p} \right)$$

Noting that $h(\cdot)$ is concave, and

$h'(p)=0$ if $p=\frac{1}{2}$, the maxima

occurs at $p=\frac{1}{2}$.

The maximum value is

$$h\left(\frac{1}{2}\right) = \frac{1}{2} \log 2 + \frac{1}{2} \log 2$$

$$= 1$$

... . 

* Hypothesis Testing - Bayesian setting

$H_0: X \sim \text{i.i.d } P_0$

$H_1: X \sim \text{i.i.d. } P_1$

Assume $P(H_0) = 1 - P(H_1) = f_0$

Hypothesis test is a map

$T: \mathcal{X} \rightarrow \{0, 1\}$ and $\Gamma = T^{-1}(1)$.

In the Bayesian setting, we look

at probability of error

$$P_e(\Gamma) = f_0 P_0(\Gamma) + (1-f_0) \cdot P_1(\Gamma^c)$$

Lemma:- Let $\Gamma^* = \{x : \frac{P_1(x)}{P_0(x)} \geq \frac{f_0}{(1-f_0)}\}$.

Then $\Gamma^* \in \operatorname{argmin}_{\Gamma'} P_e(\Gamma')$.

Proof:- Fix a $r' = \tau'^{-1}(1)$.

Then,

$$\begin{aligned} P_e(r') &= p_0 P_0(r') + (1-p_0) \cdot (1 - P_1(r')) \\ &= (p_0 P_0(r') - (1-p_0) \cdot P_1(r')) \\ &\quad + (1-p_0) \\ &= \sum_{x \in r'} (p_0 P_0(x) - (1-p_0) \cdot P_1(x)) \\ &\quad + (1-p_0) \end{aligned}$$

Note that the expression is minimised

if $\sum_{x \in r'} (p_0 P_0(x) - (1-p_0) \cdot P_1(x))$ is

minimised, i.e., if

$$r' = \{x : p_0 P_0(x) \leq (1-p_0) \cdot P_1(x)\} = r^*$$

- Theorem :- Consider the hypothesis testing setting with

$$H_0: X \sim P_0$$

$$H_1: X \sim P_1$$

with priors $P(H_0) = 1 - P(H_1) = p_0$.

Let $\Gamma^* = \{x : \frac{P_1(x)}{P_0(x)} \geq \frac{p_0}{1-p_0}\}$.

$$\begin{aligned} \text{Then, } p_0(1-p_0) \cdot 2^{-J(P_0, P_1)/2} &\leq P_E(\Gamma^*) \\ &\leq \sqrt{p_0(1-p_0)} \rho(P_0, P_1) \end{aligned}$$

where $J(P_0, P_1) \triangleq D(P_0 || P_1) + D(P_1 || P_0)$

$$J(P_0, P_1) \triangleq \sum_{x \in \mathcal{X}} \sqrt{P_0(x) P_1(x)}.$$

Remark: - $J(P_0 || P_1)$ is referred to as J-divergence, while $\rho(P_0, P_1)$ is referred to as the Bhattacharya coefficient.

Proof: Observe that

$$\begin{aligned}
 P_0(\Gamma^*) &= \sum_{x \in \Gamma^*} P_0(x) \\
 &\leq \sum_{x \in \Gamma^*} \sqrt{\frac{P_0(x)(1-f_0)}{P_0(x)f_0}} \cdot P_0(x) \\
 &\quad \left[\because x \in \Gamma^* \Rightarrow \frac{P_1(x)}{P_0(x)} \geq \frac{f_0}{(1-f_0)} \right] \\
 &= \sqrt{\frac{1-f_0}{f_0}} \sum_{x \in \Gamma^*} \sqrt{P_0(x)P_1(x)}
 \end{aligned}$$

Similarly, $P_1(\Gamma^{*c}) = \sum_{x \in \Gamma^{*c}} P_1(x)$

$$\leq \sum_{x \in \Gamma^{*c}} \sqrt{\frac{P_0(x) f_0}{P_1(x)(1-f_0)}} P_p(x)$$

$\because x \in \Gamma^{*c} \Rightarrow$

$$\frac{P_1(x)}{P_0(x)} < \frac{f_0}{1-f_0}$$

$$= \sqrt{\frac{f_0}{1-f_0}} \sum_{x \in \Gamma^{*c}} \sqrt{P_0(x) P_1(x)}$$

Then,

$$P_E(\Gamma^*) = f_0 P_0(\Gamma^*) + (1-f_0) P_1(\Gamma^*)$$

$$\leq \sqrt{f_0(1-f_0)} \left(\sum_{x \in \Gamma^*} \sqrt{P_0(x) P_1(x)} + \sum_{x \in \Gamma^{*c}} \sqrt{P_0(x) P_1(x)} \right)$$

$$\leq \sqrt{f_0(1-f_0)} \cdot p(P_0, P_1)$$

On the other hand, define

$$K(P_0, P_1) \triangleq \frac{1}{2} \sum_{x \in X} |f_{P_0} P_0(x) - (1-f_{P_0}) \cdot P_1(x)|$$

$$= \frac{1}{2} \sum_{x \in X} \left| \sqrt{f_{P_0} P_0(x)} - \sqrt{(1-f_{P_0}) P_1(x)} \right| \cdot \\ \left| \sqrt{f_{P_0} P_0(x)} + \sqrt{(1-f_{P_0}) P_1(x)} \right|$$

$$\leq \frac{1}{2} \left(\sum_{x \in X} \left(\sqrt{f_{P_0} P_0(x)} - \sqrt{(1-f_{P_0}) P_1(x)} \right)^2 \right)^{\frac{1}{2}}$$

$$\left(\sum_{x \in X} \left(\sqrt{f_{P_0} P_0(x)} + \sqrt{(1-f_{P_0}) P_1(x)} \right)^2 \right)^{\frac{1}{2}}$$

(By C-S inequality)

$$= \frac{1}{2} \left(\sum_{x \in X} \left(f_{P_0} P_0(x) + (1-f_{P_0}) P_1(x) \right. \right. \\ \left. \left. - 2\sqrt{f_{P_0} (1-f_{P_0}) P_1(x) \cdot P_0(x)} \right) \right)^{\frac{1}{2}}$$

$$\left(\sum_{x \in X} \left(f_{P_0} P_0(x) + (1-f_{P_0}) \cdot P_1(x) \right. \right. \\ \left. \left. + 2\sqrt{f_{P_0} (1-f_{P_0}) \cdot P_1(x) \cdot P_0(x)} \right) \right)^{\frac{1}{2}}$$

$$= \frac{1}{2} \sqrt{\left((1 - 2\sqrt{f_{p_0}(1-f_{p_0})} \cdot \rho(p_0, p_1)) \right) \left((1 + 2\sqrt{f_{p_0}(1-f_{p_0})} \cdot \rho(p_0, p_1)) \right)}$$

$$= \frac{1}{2} \sqrt{1 - 4f_{p_0}(1-f_{p_0}) \cdot \rho^2(p_0, p_1)}.$$

... @

Also, $P_E(\Gamma^*)$

$$= f_{p_0} P_0(\Gamma^*) + (1-f_{p_0}) P_1(\Gamma^{*c})$$

$$= f_{p_0} P_0(\Gamma^*) + (1-f_{p_0}) (1 - P_1(\Gamma^*))$$

$$= (1-f_{p_0}) + f_{p_0} P_0(\Gamma^*) - (1-f_{p_0}) \cdot P_1(\Gamma^*)$$

$$= (1-f_{p_0}) - \sum_{x \in \Gamma^*} |f_{p_0} P_0(x) - (1-f_{p_0}) \cdot P_1(x)|$$

$\left[\because x \in \Gamma^* \right]$

$$\Rightarrow P_1(x)(1-f_{p_0}) \\ \geq P_0(x) \cdot f_{p_0}$$

$$\begin{aligned}
 \text{Similarly, } P_E(\Gamma^*) &= f_0(1 - P_0(\Gamma^{*c})) \\
 &\quad + (1 - f_0) \cdot P_1(\Gamma^{*c}) \\
 &= f_0 + (1 - f_0) \cdot P_1(\Gamma^{*c}) \\
 &\quad - f_0 P_0(\Gamma^{*c}) \\
 &= f_0 - \sum_{x \in \Gamma^{*c}} |(1 - f_0) \cdot P_1(x) \\
 &\quad - f_0 P_0(x)|
 \end{aligned}$$

$\left[\because x \in \Gamma^{*c} \Rightarrow \right]$

$$f_0 P_0(x) \geq (1 - f_0) \cdot P_1(x)$$

Thus,

$$\begin{aligned}
 2P_E(\Gamma^*) &= f_0 + (1 - f_0) \\
 &\quad - \sum_{x \in \Gamma} |f_0 P_0(x) - (1 - f_0) P_1(x)| \\
 &= 1 - 2 \cdot K(P_0, P_1)
 \end{aligned}$$

Then, by Q, we have,

$$2P_E(r^*) \geq 1 - \sqrt{1 - 4f_0(1-f_0)\rho^2(P_0, P_1)}$$

$$\Rightarrow P_E(r^*) \geq \frac{1}{2} \left(1 - \sqrt{1 - 4f_0(1-f_0)\rho^2(P_0, P_1)} \right)$$

$$= \frac{1}{2} \frac{(1 - 1 + 4f_0(1-f_0)\rho^2(P_0, P_1))}{1 + \sqrt{1 - 4f_0(1-f_0)\rho^2(P_0, P_1)}}$$

$$\geq \underline{2f_0(1-f_0)\rho^2(P_0, P_1)}$$

$$= f_0(1-f_0) \rho^2(P_0, P_1).$$

Finally. $\rho^2(P_0, P_1)$

$$= 2^{\log \rho(P_0, P_1)}$$

$$= 2^{\left(\log \sum_{x \in X} \sqrt{P_0(x)P_1(x)} + \log \sum_{x \in X} \sqrt{P_0(x) \cdot P_1(x)} \right)}$$

$$= 2^{\left(\log \sum_{x \in X} \sqrt{\frac{P_0(x)}{P_1(x)}} \cdot P_1(x) + \log \sum_{x \in X} \sqrt{\frac{P_1(x)}{P_0(x)}} \cdot P_0(x) \right)}$$

$$\geq 2 \left(\sum_{x \in X} P_1(x) \cdot \log \sqrt{\frac{P_0(x)}{P_1(x)}} + \sum_{x \in X} P_0(x) \cdot \log \sqrt{\frac{P_1(x)}{P_0(x)}} \right)$$

[By Jensen's inequality
since \log is concave]

$$= 2^{-\frac{1}{2}} \left[\sum_{x \in X} P_1(x) \cdot \log \frac{P_1(x)}{P_0(x)} + \sum_{x \in X} P_0(x) \log \frac{P_0(x)}{P_1(x)} \right]$$

$$= 2^{-\frac{1}{2}} [D(P_1 || P_0) + D(P_0 || P_1)]$$

$$= 2^{-\frac{1}{2}} J(P_0 || P_1)/2$$

... ... \square

Define,

$$H(P_0, P_1) \stackrel{\Delta}{=} \sqrt{\frac{1}{2} \sum_{x \in X} (\sqrt{P_0(x)} - \sqrt{P_1(x)})^2}$$

We call $H(P_0, P_1)$ as Hellinger distance.

• Lemma :- $H(P_0, P_1)$ is a distance and

$$H(P_0, P_1) = \sqrt{1 - \rho(P_0, P_1)}.$$

Proof :- It is obvious that

$$H(P_0, P_1) \geq 0 \text{ and } H(P_0, P_1) = 0$$

iff $P_0 = P_1$, and $H(P_0, P_1) = H(P_1, P_0)$.

Now,

$$H(P_0, P_1) + H(P_1, P_2)$$

$$= \frac{1}{\sqrt{2}} \left[\sqrt{\sum_{x \in X} (\sqrt{P_0(x)} - \sqrt{P_1(x)})^2} + \sqrt{\sum_{x \in X} (\sqrt{P_1(x)} - \sqrt{P_2(x)})^2} \right]$$

$$\geq \frac{1}{\sqrt{2}} \sqrt{\sum_{x \in X} (\sqrt{P_0(x)} - \sqrt{P_1(x)} + \sqrt{P_1(x)} - \sqrt{P_2(x)})^2}$$

[By Minkowski's
inequality]

$$= H(P_0, P_2).$$

Thus, $H(\cdot, \cdot)$ is a metric.

Finally,

$$H(P_0, P_1)$$

$$= \frac{1}{\sqrt{2}} \sqrt{\sum_{x \in X} (\sqrt{P_0(x)} - \sqrt{P_1(x)})^2}$$

$$= \frac{1}{\sqrt{2}} \sqrt{\sum_{x \in X} P_0(x) + P_1(x) - 2\sqrt{P_0(x)P_1(x)}}$$

$$= \frac{1}{\sqrt{2}} \sqrt{2(1 - \rho(P_0, P_1))}$$

$$= \sqrt{1 - \rho(P_0, P_1)}$$

..... 

Remark:- Thus, as P_0 and P_1 come closer, the Bhattacharya parameter $\rho(P_0, P_1)$ increases since $H(P_0, P_1)$ decreases. Thus, the upper bound $\sqrt{f_{P_0}(1-f_{P_0})} \rho(P_0, P_1)$ increases as it should.

* f-divergence

- Definition :- Let $f: [0, \infty) \rightarrow (-\infty, \infty]$ be a convex function s.t. $f(x) \in \mathbb{R}$ for $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \rightarrow 0^+} f(t)$. Then, for any pair of pmfs P_0, P_1 on \mathcal{X} , the f-divergence of P_0, P_1 is defined as

$$D_f(P_0 || P_1) = \sum_{x \in \mathcal{X}} P_1(x) f\left(\frac{P_0(x)}{P_1(x)}\right)$$

Examples :-

i) Let $f(x) = x \log x$.

f satisfies all the required criteria.

For this f ,

$$D_f(P_0 \parallel P_1) = \sum_{x \in X} P_1(x) \cdot \frac{P_0(x)}{P_1(x)} \log \frac{P_0(x)}{P_1(x)}$$
$$= D(P_0 \parallel P_1).$$

ii) $f(x) = \frac{1}{2} |x - 1|$

Again, this f satisfies all the properties of the definition.

$$D_f(P_0 \parallel P_1) = \sum_{x \in X} P_1(x) \cdot \frac{1}{2} \left| \frac{P_0(x)}{P_1(x)} - 1 \right|$$
$$= \frac{1}{2} \sum_{x \in X} |P_0(x) - P_1(x)|$$
$$= TV(P_0, P_1)$$

iii) $f(x) = 1 - \sqrt{x}$

$$D_f(P_0 \parallel P_1) = \sum_{x \in X} P_1(x) \cdot \left(1 - \sqrt{\frac{P_0(x)}{P_1(x)}}\right)$$

$$= \sum_{x \in \mathcal{X}} \sqrt{P_1(x)} \cdot (\sqrt{P_1(x)} - \sqrt{P_0(x)})$$

$$= 1 - \sum_{x \in \mathcal{X}} \sqrt{P_1(x) \cdot P_0(x)}$$

$$= 1 - D(P_0, P_1)$$

$$= H^2(P_0, P_1)$$

$\Rightarrow f(x) = -(x+1) \log\left(\frac{x+1}{2}\right) + x \log x$

Here, it is easy to see that

$$f(x) < \infty \text{ if } x > 0$$

and $f(1) = -2 \log 1 + \log 1 = 0$

Moreover,

$$f'(x) = -2 \cdot \log e \cdot \frac{1}{2} - \log(x+1)$$

$$+ \log e + \log x = \log\left(\frac{x}{x+1}\right)$$

$$f''(x) = \frac{\log e}{x} - \frac{\log e}{x+1}$$

$$\geq 0 \quad \text{if } x \geq 0.$$

Thus $f : [0, \infty) \rightarrow (-\infty, \infty]$ is

convex.

$$D_f(P_0 || P_1) = - \sum_{x \in X} P_1(x) \left(\frac{P_0(x)}{P_1(x)} + 1 \right) \cdot \log \left(\frac{\frac{P_0(x)}{P_1(x)} + 1}{2} \right)$$

$$+ \sum_{x \in X} P_1(x) \cdot \frac{P_0(x)}{P_1(x)} \log \frac{P_0(x)}{P_1(x)}$$

$$= - \sum_{x \in X} (P_0(x) + P_1(x)) \log \left(\frac{P_0(x) + P_1(x)}{2 P_1(x)} \right)$$

$$+ \sum_{x \in X} P_0(x) \log \frac{P_0(x)}{P_1(x)}$$

$$= \sum_{x \in X} P_1(x) \cdot \log \left(\frac{2 P_1(x)}{P_0(x) + P_1(x)} \right)$$

$$\begin{aligned}
& + \sum_{x \in \Omega} P_0(x) \log \left(\frac{P_0(x)}{P_1(x)} \cdot \frac{2P_1(x)}{(P_0(x) + P_1(x))} \right) \\
& = D(P_1 \parallel \frac{P_0 + P_1}{2}) + \sum_{x \in \Omega} P_0(x) \log \frac{2P_0(x)}{(P_0(x) + P_1(x))} \\
& = D(P_1 \parallel \frac{P_0 + P_1}{2}) + D(P_0 \parallel \frac{P_0 + P_1}{2}) \\
& \triangleq JS(P_0, P_1)
\end{aligned}$$

$JS(P_0, P_1)$ is also referred to as
Jensen-Shannon divergence.

• Lemma :- $\mathcal{D}_{f_1+f_2} = \mathcal{D}_{f_1} + \mathcal{D}_{f_2}$.

- i) $\mathcal{D}_f = 0 \iff f(x) = c(x-1)$
- ii) $\mathcal{D}_{f+g} = \mathcal{D}_f \iff g(x) = c(x-1)$.

Proof:- ii) directly follows from i)
and ii). We only need to

prove (i) and (ii).

(i) Firstly note that if f_1 and f_2 satisfies the properties needed to define an f -divergence, then so does $f_1 + f_2$. Thus, $D_{f_1 + f_2}$ is well-defined.

Now,

$$\begin{aligned} D_{f_1 + f_2}(P_0 \parallel P_1) &= \sum_{x \in X} P_1(x) \left(f_1\left(\frac{P_0(x)}{P_1(x)}\right) \right. \\ &\quad \left. + f_2\left(\frac{P_0(x)}{P_1(x)}\right) \right) \\ &= D_{f_1}(P_0 \parallel P_1) + D_{f_2}(P_0 \parallel P_1). \end{aligned}$$

ii) Let $f = c(x-1)$.

Then, for any P_0, P_1 ,

$$\begin{aligned} D_f(P_0 \parallel P_1) &= \sum_{x \in X} P_1(x) \cdot c \left(\frac{P_0(x)}{P_1(x)} - 1 \right) \\ &= c \cdot \sum_{x \in X} (P_0(x) - P_1(x)) = 0. \end{aligned}$$

On the other hand let there exist some f which satisfy the properties needed by f -divergences s.t. $D_f = 0$.

Choose $I = \{0, 1\}$ and two pmfs

$$P_0 = Be(f), \quad P_1 = Be(g), \quad f > g.$$

$$D_f(P_0 || P_1) = qf\left(\frac{f}{g}\right) + (1-q)f\left(\frac{1-f}{1-g}\right)$$

$$\Rightarrow \frac{f\left(\frac{f}{g}\right)}{f\left(\frac{1-f}{1-g}\right)} = - \frac{1-q}{q} \quad \dots \quad (A)$$

Let $\frac{f}{g} = a$, $\frac{1-f}{1-g} = b$. Note that as

$$f > g, \quad a > 1 > b > 0.$$

Now, plugging $f_i = aq$, we get

$$(a-b) \cdot q = 1-b \\ \Rightarrow q = \frac{1-b}{a-b}$$

Then, plugging this in ④ we get

$$\frac{f(a)}{f(b)} = - \frac{(a-1)}{(1-b)}, \dots \textcircled{B}$$

Now consider any $x \in [1, a]$, and note that x can be written as

$$x = \lambda a + (1-\lambda) b$$

By convexity of f , we have

$$f(x) \leq \lambda f(a) \quad [\because f'(1)=0]$$

Assume $f(x) < \lambda f(a)$.

Next, let $\mu = \frac{1-b}{x-b}$, and note that

$$0 < \mu < 1.$$

$$\text{Also, } l = \mu x + (1-\mu)b$$

Thus, by convexity of f and $f(l)=0$,

$$\mu f(x) + (1-\mu)f(b) \geq 0$$

$$\Rightarrow \mu \lambda f(a) + (1-\mu)f(b) \geq 0$$

$$\Rightarrow \frac{(1-b)}{(x-b)} \cdot \frac{(x-1)}{(a-1)} f(a) + \frac{(x-1)}{(x-b)} f(b) \geq 0$$

$$\Rightarrow (1-b)f(a) + (a-1)f(b) \geq 0,$$

which violates (B).

Hence, our assumption $f(w) < \lambda f(a)$ was wrong, and we must have

$$f(x) = \lambda f(a), \text{ i.e. } f(x) = \frac{(x-1)}{(a-1)} \cdot f(a).$$

Noting that the choice of $x \in [1, a]$ was arbitrary, we have f to be

linear in $[1, a]$ with slope $\frac{f(a)}{a-1}$

Similarly, we can show that

f is linear in $[b, 1]$ with slope

$$-\frac{f(b)}{(1-b)} = \frac{f(a)}{a-1} \quad (\text{by (B)})$$

Thus, f is linear in $[a, b]$.

Now, note that the choice of a and b were arbitrary, and one can choose a to be arbitrarily large and b to be arbitrarily small by choosing sufficiently small q and sufficiently large p respectively.

Thus, f is linear in $[0, \infty)$.

Thus, since $f(1) = 0$, f must be of the form $f(x) = c(x-1)$.
... \square

• Lemma :-

(i) $D_f(P_0 \parallel P_1) \geq 0$. If f is strictly convex, then $D_f(P_0 \parallel P_1) = 0$ iff $P_0 = P_1$.

(ii) The map $(P_0, P_1) \rightarrow D_f(P_0 \parallel P_1)$ is convex.

Proof:- ⁽ⁱ⁾ $D_f(P_0 \parallel P_1)$

$$= \sum_{x \in X} P_1(x) f\left(\frac{P_0(x)}{P_1(x)}\right)$$

$$= E_{P_1} \left[f\left(\frac{P_0(x)}{P_1(x)}\right) \right]$$

$$\geq f\left(E_{P_1}\left(\frac{P_0(x)}{P_1(x)}\right)\right) \quad [\text{Jensen, as } f \text{ is convex}]$$

$$= f \left(\sum_{x \in X} P_0(x) \cdot \frac{P_0(x)}{P_1(x)} \right)$$

$$= f(1)$$

$$= 0 \quad [\because \beta_f \text{ defn of } f]$$

If f is strictly convex, then we

have equality iff $\frac{P_0(x)}{P_1(x)} = \text{constant}$,

$$\text{i.e. } P_0(x) = P_1(x).$$

(ii) Fix any $\lambda \in [0, 1]$ and let

$P_0^{(1)}, P_1^{(1)}, P_0^{(2)}, P_1^{(2)}$ be pmfs on X .

Then,

$$\lambda D_f(P_0^{(1)} \| P_1^{(1)}) + (1-\lambda) \cdot D_f(P_0^{(2)} \| P_1^{(2)})$$

$$= \sum_{x \in X} \left(\lambda P_0^{(1)}(x) f\left(\frac{P_0^{(1)}(x)}{P_1^{(1)}(x)}\right) + (1-\lambda) \cdot P_1^{(2)}(x) f\left(\frac{P_0^{(2)}(x)}{P_1^{(2)}(x)}\right) \right)$$

$$= \sum_{x \in X} (\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)) \mathbb{E}[f(Y_x)]$$

where $P(Y_x = \frac{P_0^{(1)}(x)}{P_1^{(1)}(x)}) = \frac{\lambda P_1^{(1)}(x)}{\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)}$

$$P(Y_x = \frac{P_0^{(2)}(x)}{P_1^{(2)}(x)}) = \frac{(1-\lambda) P_1^{(2)}(x)}{\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)}$$

$$\geq \sum_{x \in X} (\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)) f(\mathbb{E}[Y_x])$$

[Jensen, as f is convex]

$$= \sum_{x \in X} (\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)) f\left(\frac{\lambda P_0^{(1)}(x) + (1-\lambda) P_0^{(2)}(x)}{\lambda P_1^{(1)}(x) + (1-\lambda) P_1^{(2)}(x)}\right)$$

$$= D_f(\lambda P_0^{(1)} + (1-\lambda) P_0^{(2)} || \lambda P_1^{(1)} + (1-\lambda) P_1^{(2)})$$

. . .

* Variable length source coding

- Motivation

- We know that the optimal source code has compression rate $H(x)$.
- We don't know what that code is.
- Can we reach close to $H(x)$?

• Definition :- Given a discrete set \mathcal{X}

a variable length source code constitutes of two maps

$$l: \mathcal{X} \rightarrow \mathbb{N} \text{ and } C: \mathcal{X} \rightarrow \{0,1\}^*$$

$$\text{s.t. } C(x) \in \{0,1\}^{l(x)}.$$

The map l is called the length function, and the map C is called the encoder.

Issue:- How does one decode a variable length code?

e.g. Let $\mathcal{X} = \{a, b, c\}$

Define $l(a) = 2$, $l(b) = 2$, $l(c) = 1$

$C(a) = 10$, $C(b) = 11$, $C(c) = 1$

Suppose we see the encoded string

111

Note that we can interpret this string as

1, 1, 1 i.e. $C(c) C(c) C(c)$

11, 1 i.e. $C(b) \cdot C(c)$

1, 11 i.e. $C(c) \cdot C(b)$

In other words, every variable length source code is not uniquely decodable.

- Definition :- A variable length source code C is uniquely decodable if for any $k, l \in \mathbb{N}$ and any choice of $x_{1_1}, x_{1_2}, \dots, x_{1_k} \in \mathcal{X}$ and $x_{2_1}, x_{2_2}, \dots, x_{2_l} \in \mathcal{X}$,
 $C(x_{1_1})C(x_{1_2}) \dots C(x_{1_k}) \neq C(x_{2_1})C(x_{2_2}) \dots C(x_{2_l})$.

- Definition' :- A variable length source code $C : \mathcal{X} \rightarrow \{0,1\}^*$ is called prefix free, if $\forall x \in \mathcal{X}, \exists x' \neq x$ s.t.

$$C(x) = C(x_1) \circ y, \text{ for some } y \in \{0,1\}^*$$

• Proposition:- Every prefix free code is uniquely decodable.

Proof: We will show that every code which is not uniquely decodable must not be prefix free.

To see this assume $C: X \rightarrow \{0,1\}^*$ is not uniquely decodable. Then,

$$\exists x_{1,1}, \dots, x_{1,k}, x_{2,1}, \dots, x_{2,l} \in X \text{ s.t}$$

$$C(x_{1,1}) \circ C(x_{1,2}) \circ \dots \circ C(x_{1,k})$$

$$= C(x_{2,1}) \circ C(x_{2,2}) \circ \dots \circ C(x_{2,l}).$$

Then, we must have that $\exists 1 \leq j \leq \min\{k, l\}$

s.t. $C(x_{1:i}) = C(x_{1:j})$, $\forall 1 \leq j \leq i-1$,

and $C(x_{1:j}) = C(x_{2:j}) \circ y$

or $C(x_{2:j}) = C(x_{1:j}) \circ y$, for some $y \in \{0, 1\}^*$.

Hence, C is not prefix-free.

□

• Lemma (Kraft inequality) :-

Let X be a discrete set and let

$l: X \rightarrow \mathbb{N}$ be any function. Then

l is a length function for a

prefix free code iff $\sum_{x \in X} q^{-l(x)} \leq 1$.

Proof: First we assume that l is indeed the length function of some prefix free code c . We construct a (possibly incomplete) binary tree inductively as follows.

Suppose we are at a node v in the tree. Let $y \in \{0,1\}^*$ be the string of bits that represent the unique path from root to v , and let y be the prefix to codewords $((x_1), \dots, (x_n))$.

If $y00$ is also a prefix to some of these codewords, add the 0-child node of v . If $y01$ is also a prefix to at least one of $((x_1), \dots, (x_n))$, add the 1-child node of v .

We say that node v represents some codeword $C(x)$, if the path of root to v is $C(x)$.

Since C is prefix-free, a node represents a codeword $C(x)$ if and only if v is a leaf.

Now, let $l_{\max} = \max_{x \in I} l(x)$, and note that the depth of the tree is l_{\max} .

Next, consider a leaf at depth l' . Then

(i) If the tree was extended to a complete binary tree of depth l_{\max} , the number of leaves of the complete binary tree contributed by this leaf is $2^{l_{\max}-l'}$.

(ii) No other leaf of the original tree will

contribute to these leaves of the complete binary tree.

Thus, noting that a complete binary tree of depth l_{\max} can have at most $2^{l_{\max}}$ leaves, we have

$$\sum_{x \in \Sigma} 2^{l_{\max} - l(x)} \leq 2^{l_{\max}}$$

$$\Rightarrow \sum_{x \in \Sigma} 2^{-l(x)} \leq 1.$$

To prove the opposite direction, let $\ell: \Sigma \rightarrow \mathbb{N}$ be a length function satisfying

$$\sum_{x \in \Sigma} 2^{-\ell(x)} \leq 1.$$

Now take a complete binary tree of depth $l_{\max} = \max_{x \in X} l(x)$.

We shall prune this tree as follows.

- Label the elements of set X as

$$X = \{1, 2, \dots, |X|\}$$

- Consider the first $2^{l_{\max} - l(1)}$ leaves of the complete binary tree from the left.

Prune the tree to the least common ancestor of these leaves. Note that

this ancestor is at a depth $l(1)$ since we chose $2^{l_{\max} - l(1)}$ consecutive leaves.

Further note that since we chose a number of leaves which is a power of 2, no extra leaf is taken out when we prune the tree to the least common ancestor of these leaves.

- We continue this pruning process for the next consecutive $2^{\ell_{\max} - \ell(2)}, 2^{\ell_{\max} - \ell(3)}, \dots$ $2^{\ell_{\max} - \ell(\lceil x \rceil)}$ leaves, to their respective ancestors at levels $\ell(2), \ell(3), \dots, \ell(\lceil x \rceil)$.

Note that this is possible because

$$(i) \sum_{x \in X} 2^{-\ell(x)} \leq 1 \quad \text{and}$$

(ii) No 'extra' leaf is taken out during pruning.

Now consider the pruned binary tree with leaves at levels $\ell(x), x \in X$.

Define the code $C: X \rightarrow \{0,1\}^*$ as

$C(x)$ = unique path from root to the leaf at level $\ell(x)$.

C is obviously prefix free since it reads paths from root to leaves of the pruned tree

□

* Metrics for prefix-free code

• Lemma: - let X be a source equipped with a pmf P_X . Then

(i) Any prefix free code $C: X \rightarrow \{0,1\}^*$

satisfies $\mathbb{E}[l(x)] \geq H(P_X)$.

(ii) There exists a prefix-free code

$C_{SF}: X \rightarrow \{0,1\}^*$ satisfying

$$\mathbb{E}[l_{SF}(x)] \leq H(P_X) + 1.$$

Proof: (i) $E[l(x)] = H(p_x)$

$$= \sum_{x \in X} p_x(x) \cdot \log_2 e^{l(x)} + \sum_{x \in X} p_x(x) \cdot \log_2 p_x(x)$$

$$\geq \sum_{x \in X} p_x(x) \log_2 \frac{p_x(x)}{2^{-l(x)}}$$

$$\geq \left(\sum_{x \in X} p_x(x) \right) \cdot \log_2 \frac{\left(\sum_{x \in X} p_x(x) \right)}{\left(\sum_{x \in X} 2^{-l(x)} \right)}$$

$$= \log_2 \frac{1}{\left(\sum_{x \in X} 2^{-l(x)} \right)}$$

$$\geq \log_2 1 \quad [\text{Kraft inequality and } \log \text{ is increasing}]$$

$$= 0$$

(ii) Consider the length function

$$l_{SF}(x) = \lceil \log \frac{1}{P_X(x)} \rceil$$

Then,

$$\sum_{x \in X} 2^{-l_{SF}(x)} = \sum_{x \in X} 2^{-\lceil \log \frac{1}{P_X(x)} \rceil}$$

$$\leq \sum_{x \in X} 2^{-\log \frac{1}{P_X(x)}}$$

$$= \sum_{x \in X} P_X(x) = 1$$

Thus, by Kraft inequality, there exists a prefix free code $G_F: X \rightarrow \{0,1\}^*$, whose length function is $l_{SF}(x) = \lceil \log \frac{1}{P_X(x)} \rceil$.

Then,

$$\begin{aligned} \mathbb{E}[l_{SF}(x)] &= \sum_{x \in X} P_X(x) \lceil \log \frac{1}{P_X(x)} \rceil \\ &\leq \sum_{x \in X} P_X(x) \left(\log \frac{1}{P_X(x)} + 1 \right) \\ &= H(P_X) + 1 \dots \square \end{aligned}$$

Remark:- The prefix free code ($s_F : X \rightarrow \{0,1\}^*$)
is known as the Shannon - Fano code.

Example:-

$$\text{Let } X = \{1, 2, 3, 4, 5, 6\}$$

$$P_X = (0.3, 0.25, 0.2, 0.12, 0.08, 0.05)$$

$$\left\lceil \log \frac{1}{P_X} \right\rceil = (2, 2, 3, 4, 4, 5)$$

$$l_{\max} = 5$$

$$C(1) = 00$$

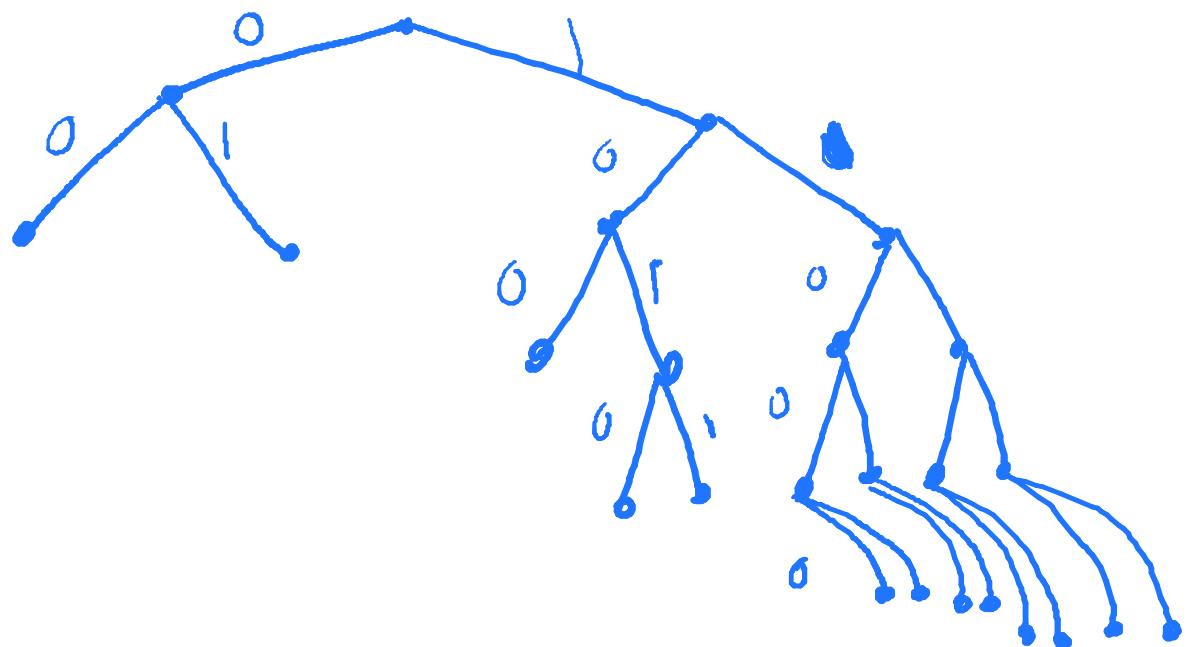
$$C(2) = 01$$

$$C(3) = 100$$

$$C(4) = 1010$$

$$C(5) = 1011$$

$$C(6) = 1100 \text{ D}$$



* Optimal variable length codes

— Huffman codes.

Huffman codes are constructed as follows.

Input : Source X and pmf $P_X(i) = p_i$

— Relabel X s.t.

$$p_1 \geq p_2 \geq \dots \geq p_{|X|}$$

— If $|X|=2$, use $C_H^{(1)}=0$, $C_H^{(2)}=1$.

- Else
 - Define new source
 $X' = \{1', 2', \dots, (x_{l-1})'\}$
 new pmf $P_{X'} \text{ s.t.}$
 - $P_{X'}(i') = f_i, \forall 1 \leq i' \leq |X|-2$
 - $P_{X'}((x_{l-1})') = f_{|X|} + f_{|X|-1}$
 - Apply Huffman coding on this new source to obtain $C_H^{(x_{l-1})}$.
 - Define $C_H^{(x_l)}$ as follows.

$$C_H^{(x_l)}(i) = C_H^{(x_{l-1})}(i'), \quad \forall 1 \leq i \leq |X|-2$$

$$C_H^{(x_l)}(|X|-1) = C_H^{(x_{l-1})}((x_{l-1})') \neq 0$$

$$C_H^{(x_l)}(|X|) = C_H^{(x_{l-1})}((x_{l-1})') \neq 0$$

• Example:-

x	P_x	
1	0.3	0.3 ... 0.3 ... 0.3 ... 0.55
2	0.25	0.25 ... 0.25 ... 0.25 ... 0
3	0.2	0.2 ... 0.45 ... 0.45
4	0.12	0.12 ... 0.25 ... 0.25
5	0.08	0.13 ... 0.13 ... 0.25
6	0.05	0

$$C_H(1) = 11$$

$$C_H(2) = 10$$

$$C_H(3) = 00$$

$$C_H(4) = 010$$

$$C_H(5) = 0111$$

$$C_H(6) = 0110$$

• Definition:- Consider a source with a relabelled alphabet X s.t.

$$P_X(i) = p_i \text{ and } p_1 \geq p_2 \geq \dots \geq p_{|X|}.$$

A variable length code $C: X \rightarrow \{0,1\}^*$

is called canonical if

$$(i) \quad l(i) \leq l(j), \quad \forall i \leq j$$

(ii) $l_{max} = C(1x1) = C(1x1-1)$, and
 $C(1x1)$ and $C(1x1-1)$ are 'siblings', i.e.,

$\exists y \in \{0,1\}^{l_{\max}-1}$ s.t.

$$C(|x|-1) = y \cdot 0 \quad \text{or} \quad C(|x|-1) = y \cdot 1$$

$$C(|x|) = y \cdot 1 \quad C(|x|) = y \cdot 0.$$

• Definition:- We call a code $C: X \rightarrow \{0,1\}^*$

optimal if for any other code

$C': X \rightarrow \{0,1\}^*$, we have

$$\mathbb{E}_{P_x}[\ell(x)] \leq \mathbb{E}_{P_x}[\ell'(x)],$$

where ℓ, ℓ' are respectively the length functions of C, C' .

• Lemma:- Any source with alphabet \mathcal{N} and pmf P_x has an optimal code which is canonical.

Proof: Let $C^*: \mathcal{X} \rightarrow \{0,1\}^*$ be an optimal code.

We assume that the source is relabelled to ensure $f_1 \geq f_2 \geq \dots$.

Now suppose $\exists i \leq j$ s.t. $\hat{l}(i) > \hat{l}(j)$.

Define a new code C' where

$$\begin{aligned} C'(k) &= \hat{l}(k) \text{ if } k \neq i, j \\ &= C^*(i) \text{ if } k = j \\ &= C^*(j) \text{ if } k = i \end{aligned}$$

Observe that C' is also prefix free.

$$\begin{aligned} \text{But } E_{P_x}[l'(x)] &= \sum_{k \neq i, j} l'(k) f_k + f_i l'(i) \\ &\quad + f_j l'(j) \\ &= \sum_{k \neq i, j} \hat{l}(k) f_k + f_i \hat{l}(j) \\ &\quad + f_j \hat{l}(i) \end{aligned}$$

$$= E_{P_x} [\ell^*(x)] + f_i (\ell^*(j) - \ell^*(i)) \\ + f_j (\ell^*(i) - \ell^*(j))$$

$$= E_{P_x} [\ell^*(x)] + (f_i - f_j) (\ell^*(j) - \ell^*(i))$$

$$< E_{P_x} [\ell^*(x)],$$

which contradicts the optimality
of C^* .

Thus, we must have $\ell^*(i) \leq \ell^*(j)$
whenever $i \leq j$.

In particular, this implies

$$\ell_{\max} = \ell^*(1|x_1)$$

Next, we assume that $C^*(1|x_1)$ does
not have any $C^*(x)$, $x \neq 1|x_1$
as a sibling.

Then, we keep on deleting the final bit of $C^*(|x|)$ till we reach a point where $\exists x \neq |x|$ s.t. $C^*(x)$ is a sibling of the string obtained by bit-deletions. Then, we can define a new prefix free code where $C^*(|x|)$ is replaced by the bit deleted string and this new code will have strictly lower expected length which violates the optimality of C^* .

Thus, $\exists x' \neq |x|$ s.t. $C^*(x)$ is a sibling of $C^*(|x|)$.

Next, since $l(i) \leq l(j)$ if $i \leq j$, we must

have $C^*(x') = C^*(x'+1) = \dots = C^*(|x|-1) = C^*(|x|)$.

Now, if $x' = |x|-1$, we already have
that C^* is canonical.

If not, define a new code $C_c : X \rightarrow \{0,1\}^*$
as follows:

$$\begin{aligned}C_c(x) &= C^*(x), \text{ if } x \neq x', |x|-1 \\&= C^*(x'), \text{ if } x = |x|-1 \\&= C^*(|x|-1), \text{ if } x = x'.\end{aligned}$$

Note that by definition C_c is

prefix free and $\mathbb{E}_{P_X}[l_c(x)] = \mathbb{E}_{P_X}[l^*(x)]$.

Hence, C_c is also an optimal code.

Finally, by construction, $C_c(|x|)$ and

$C_c(|x|-1)$ are siblings, and hence

C_c is canonical.

Theorem: - For any source X and a pmf P_X , the Huffman code C_H is optimal.

Proof: - We shall prove this by induction on $|X|$.

First assume that $X = \{1, 2\}$, i.e., $|X|=2$.

Then $C_H(1) = 0$, $C_H(2) = 1$, i.e.

$$E_{P_X}[l_H(x)] = 1$$

Since for any code $l(x) \geq 1$,

$$E_{P_X}[l(x)] \geq 1.$$

Thus, C_H is optimal for $|X|=2$.

Next, assume that C_H is optimal for $|X|=n$.

Now, WLOG, let $X = \{1, 2, \dots, n+1\}$

s.t. $P_X(1) \geq P_X(2) \geq \dots \geq P_X(n+1)$.

Define a new alphabet

$X' = \{1', 2', \dots, n'\}$ with pmf $P_{X'}$:

s.t. $P_{X'}(i') = P_{X,i}(i')$, $\forall 1 \leq i \leq n-1$

and $P_{X'}(n') = P_X(n) + P_X(n+1)$.

Let C_c^* be the canonical optimal code for X and let $C_H^{(n')}$ be the Huffman code for X' .

From C_c^* we now construct a prefix-free code on X' , $C': X' \rightarrow \{0, 1\}^*$ as follows

$C'(i') = C_c^*(i)$, $\forall 1 \leq i \leq n-1$

$C'(n') = \text{Common parent of } C_c^*(n) \text{ and } C_c^*(n+1)$.

Note that $C'(n')$ is defined since C_c^* is canonical.

Also, let $C_H^{(n)}$ be the Huffman code on \mathcal{X} . Then, by definition of Huffman codes, we have

$$C_H^{(n+1)}(i) = C_H^{(n)}(i') \quad \forall 1 \leq i \leq n-1$$

$$C_H^{(n+1)}(n) = C_H^{(n)}(n') \circ 0$$

$$C_H^{(n+1)}(n+1) = C_H^{(n)}(n') \circ 1$$

Now, we calculate the expected length

of $C_H^{(n+1)}$, $C_H^{(n)}$, C' , C_c^* .

$$\begin{aligned} E_{P_{X'}}[l'(x)] &= \sum_{i=1}^n l'(i') \cdot P_{X'}(i') \\ &= \sum_{i=1}^{n-1} l_c^*(i) \cdot P_X(i) \\ &\quad + (l_c^*(n+1) - 1) \cdot (P_X(n) + P_X(n+1)) \\ &= E_{P_X}[l_{C_c^*}(x)] - (P_X(n) + P_X(n+1)) \\ &\quad [\because l_c^*(n+1) = l_c^*(n) \\ &\quad \text{as } C_c^* \text{ is ... (1)} \\ &\quad \text{canonical}] \end{aligned}$$

Above,

$$\begin{aligned}
 \mathbb{E}_{P_X} [l_H^{(n+1)}(x)] &= \sum_{i=1}^{n+1} l_H^{(n+1)}(i) \cdot P_X(i) \\
 &= \sum_{i=1}^{n+1} l_H^{(n)}(i) \cdot P_{X'}(i) \\
 &\quad + (l_H^{(n)}(n) + 1) \cdot P_X(n) \\
 &\quad + (l_H^{(n)}(n) + 1) \cdot P_X(n+1) \\
 \\
 &= \sum_{i=1}^n l_H^{(n)}(i) \cdot P_{X'}(i) \\
 &\quad + (P_X(n) + P_X(n+1))
 \end{aligned}$$

... ②

Combining ① and ②, we have

$$\begin{aligned}
 \mathbb{E}_{P_X} [l_c^*(x)] - \mathbb{E}_{P_{X'}} [l'(x)] &= \mathbb{E}_{P_X} [l_H^{(n+1)}(x)] \\
 &\quad - \mathbb{E}_{P_{X'}} [l_H^{(n)}(x)] \\
 \Rightarrow \underbrace{\left(\mathbb{E}_{P_X} [l_c^*(x)] - \mathbb{E}_{P_X} [l_H^{(n+1)}(x)] \right)}_A \\
 &\quad + \underbrace{\left(\mathbb{E}_{P_{X'}} [l_H^{(n)}(x)] - \mathbb{E}_{P_{X'}} [l'(x)] \right)}_B = 0
 \end{aligned}$$

Now, the optimality of C_c^* implies

$$A \geq 0.$$

Also, $C_H^{(m)}$ is optimal for x' by
the induction hypothesis since $|x'| = n$

$$\text{Thus } B \geq 0.$$

Then, as $A + B = 0$, we must have

$$A = 0 \text{ and } B = 0.$$

In particular, $A = 0$ implies

$$\mathbb{E}_{P_x} [l_c^*(x)] = \mathbb{E}_{P_x} [l_H^{(n+1)}(x)],$$

and since C_c^* was optimal, we
must have that $C_H^{(n+1)}$ is also
optimal.

... 

* Types and Type classes

Consider a source with alphabet \mathcal{X} .

Let $x^n \in \mathcal{X}^n$.

Define

$$N(a|x^n) = \sum_{i=1}^n 1\{\bar{x}_i = a\},$$
$$\forall a \in \mathcal{X}.$$

i.e., $N(a|x^n)$ counts the number of times a occurs in x^n .

Then, $P_{x^n}(a) = \frac{N(a|x^n)}{n}, \quad \forall a \in \mathcal{X}$

defines a pmf on \mathcal{X} .

The pmf P_{x^n} is referred to as the type of x^n .

Now, let P be a pmf on X .

Then, the type class of P is defined

as

$$T_P \stackrel{(n)}{=} \{x^n : P_{x^n} = P\}.$$

Note that there exists P such that $T_P \stackrel{(n)}{=} \emptyset$.

Now, denote the set of all types as

$$\mathcal{P}^{(n)} = \{P_{x^n} : x^n \in X^n\}.$$

• Lemma: $|P^{(n)}| \leq (n+1)^{|X|}$.

Proof:- Note that for every $a \in X$,
the number of possible values
of $N(a|x^n)$ is $(n+1)$.

Thus, $|\{P_{x^n} : x^n \in X^n\}| \leq (n+1)^{|X|}$.

• Lemma :- Let $P \in \mathcal{P}^{(m)}$. Then,

$$(M+1)^{-|X|} 2^{-nH(P)} \leq |T_P^{(n)}| \leq 2^{-nH(P)}.$$

Proof :-

$$P(x^n) = \prod_{a \in X} P(a)^{N(a/x^n)}$$

$$= \prod_{a \in X} P(a)^{n \cdot P_{x^n}(a)}$$

Now, if $x^n \in T_P^{(n)}$, then $P = P_{x^n}$.

$$\text{Hence, } P(x^n) = \prod_{a \in X} P(a)^n P(a)$$

$$= \prod_{a \in X} 2^{n P(a) \log P(a)}$$

$$= 2^{-n \sum_{a \in X} P(a) \log \frac{1}{P(a)}}$$

$$= 2^{-nH(P)}.$$

$$\text{So, } P(T_P^{(n)}) \leq 1$$

$$\Rightarrow \sum_{x^n \in T_P^{(n)}} P(x^n) \leq 1$$

$$\Rightarrow \sum_{x^n \in T_P^{(n)}} 2^{-nH(P)} \leq 1$$

$$\Rightarrow |T_P^{(n)}| \cdot 2^{-nH(P)} \leq 1$$

$$\Rightarrow |T_P^{(n)}| \leq 2^{nH(P)}$$

On the other hand, let $\hat{P} \in \Omega^{(n)}$, be some other type,

let $y^n \in T_{\hat{P}}^{(n)}$.

$$\text{Then, } P(y^n) = \prod_{\alpha \in \mathcal{X}} P(\alpha)^{n \hat{P}(\alpha)}$$

Then, $P(T_{\hat{P}}^{(n)})$

$$= \sum_{y^n \in T_{\hat{P}}^{(n)}} P(y^n)$$

$$= \sum_{y^n \in T_{\hat{P}}^{(n)}} \left(\prod_{a \in X} P(a)^{n \cdot \hat{P}(a)} \right)$$

$$= |T_{\hat{P}}^{(n)}| \cdot \left(\prod_{a \in X} P(a)^{n \cdot \hat{P}(a)} \right)$$

$$= \frac{n!}{\prod_{a \in X} (n \hat{P}(a))!} \prod_{a \in X} P(a)^{n \cdot \hat{P}(a)}$$

$$\text{Then, } \frac{P(T_{\hat{P}}^{(n)})}{P(T_P^{(n)})} = \frac{\left(\prod_{a \in X} (n \hat{P}(a))! \right)}{\left(\prod_{a \in X} (n P(a))! \right)} \prod_{a \in X} P(a)^{n(P(a) - \hat{P}(a))}$$

$$= \prod_{a \in X} \left(\frac{(n \hat{P}(a))!}{(n P(a))!} P(a)^{n(P(a) - \hat{P}(a))} \right)$$

$$\geq \prod_{a \in X} \left(n^{P(a)} \cdot P(a)^{n(\hat{P}(a) - P(a))} \right)$$

$\left[\because \frac{f!}{g!} \geq q^{f-g} \right]$

To see this consider $f > g$

then $\frac{f!}{g!} = f(f-1)\dots(g+1)$
 $\geq q^{f-g}$

If $f \leq g$, then

$$\begin{aligned} \frac{f!}{g!} &= \frac{1}{g(g-1)\dots(f+1)} \\ &\geq \frac{1}{q^{g-f}} = q^{f-g} \end{aligned}$$

$$= \prod_{a \in X} n^{n(\hat{P}(a) - P(a))}$$

$$= n^n \sum_{a \in X} (\hat{P}(a) - P(a))$$

$$= 1$$

Thus, we have proved that for

any $\hat{P} \in \mathcal{P}^{(n)}$ we have

have, $P(T_P^{(n)}) \geq P(T_{\hat{P}}^{(n)})$.

Now, observe that

$$\bigcup_{\hat{P} \in \mathcal{P}^{(n)}} T_{\hat{P}}^{(n)} = \Omega^n$$

and $T_{\hat{P}_1}^{(n)} \cap T_{\hat{P}_2}^{(n)} = \emptyset$, if $\hat{P}_1 \neq \hat{P}_2$.

Then $I = P(\Omega^n)$

$$= \sum_{\hat{P} \in \mathcal{P}^{(n)}} P(T_{\hat{P}}^{(n)})$$

$$\leq \sum_{\hat{P} \in \mathcal{P}^{(n)}} P(T_P^{(n)}) \quad (\because P(T_P^{(n)}) \geq P(T_{\hat{P}}^{(n)}))$$

$$= |\mathcal{P}^{(n)}| \cdot P(T_P^{(n)})$$

$$\leq (n+1)^{|x|} P(T_P^{(n)})$$

$$= (n+1)^{|x|} \cdot |T_p^{(n)}| \cdot 2^{-nH(p)}$$

Thus, $|T_p^{(n)}| \geq 2^{nH(p)} \cdot (n+1)^{-|x|}$

B

Lemma :- Let Q be any fmf on X . Then,

(i) If $x^n \in T_p^{(n)}$, then

$$Q(x^n) = 2^{-n(D(p||Q) + H(p))}$$

(ii) For any $P \in \mathcal{P}^{(n)}$,

$$2^{-nD(p||Q)} (n+1)^{-|X|} \leq Q(T_p^{(n)}) \leq 2^{-nD(p||Q)}$$

Proof:- (i) $Q(x^n) = \prod_{a \in x} Q(a)^{N(a|x^n)}$

$$= \prod_{a \in x} Q(a)^{n P(a)} \quad [\because x^n \in T_p^{(n)}]$$

$$\begin{aligned}
&= \prod_{a \in X} 2^{-n P(a) \log Q(a)} \\
&= \prod_{a \in X} 2^{-n \left(P(a) \log \frac{P(a)}{Q(a)} + P(a) \log \frac{1}{P(a)} \right)} \\
&= 2^{-n \sum_{a \in X} \left(P(a) \log \frac{P(a)}{Q(a)} + P(a) \log \frac{1}{P(a)} \right)} \\
&= 2^{-n(D(P||Q) + H(P))}
\end{aligned}$$

Now, $Q(T_P^{(n)})$

$$\begin{aligned}
&= \sum_{y^n \in T_P^{(n)}} Q(y^n) \\
&= \sum_{y^n \in T_P^{(n)}} 2^{-n(D(P||Q) + H(P))} \\
&= 2^{-n(D(P||Q) + H(P))} \cdot |T_P^{(n)}|
\end{aligned}$$

$$\text{Since, } (n+1)^{-|X|} 2^{-nH(P)} \leq |T_P^{(n)}| \leq 2^{nH(P)},$$

We have

$$(n+1)^{-1} 2^{-n D(P||Q)} \leq Q(T_P^{(n)}) \leq 2^{-n D(P||Q)}$$

∴ P

* Universal source coding

Theorem :- Consider a source with alphabet \mathcal{X} and pmf P_X , and any $R > 0$. there exists a sequence of encoder-decoder pairs $(f^{(n)}, g^{(n)})_{n \geq 1}$, with $f^{(n)}: \mathcal{X}^n \rightarrow \{0,1\}^{k(n)}$ and $g^{(n)}: \{0,1\}^{k(n)} \rightarrow \mathcal{X}^n$, which do not depend on P_X such that

$$(i) P_X(X^n \neq g^{(n)}(f^{(n)}(X^n))) \leq 2^{-n C^{(n)}}$$

$$(ii) \lim_{n \rightarrow \infty} \frac{k(n)}{n} = R,$$

where $C^{(n)} = \inf_{Q: H(Q) > R} D(Q || P_x) = \frac{\log(n+1)}{n} / |X|$

Remarks:-

(i) This sequence of encoder-decoder pair is called universal since they do not depend on P_x .

(ii) If the set $\{Q : H(Q) > R\} = \emptyset$, then we set $\inf_{Q: H(Q) > R} D(Q || P) = \infty$.

(iii) Note that this theorem does not violate the lossless source coding theorem. This is because if $R < H(P_x)$,

$P_x \notin \{Q : H(Q) > R\}$ and hence

$$\inf_{Q: H(Q) > R} D(Q || P_x) = D(P_x || P_x) = 0$$

$$\text{Thus, } C^{(n)} = -\frac{\log(n+1)/x_1}{n}$$

$$\Rightarrow nC^{(n)} = \log((n+1)/x_1)$$

and hence the upper bound on the probability of error given by the theorem grows to ∞ as $n \rightarrow \infty$, meaning that this scheme is unable to give a coding scheme of rate R if $H(P_x) < R$.

Proof of theorem :-

Consider the set

$$A^{(n)} \triangleq \bigcup_{Q \in \Omega^{(n)}} T_Q^{(n)}$$

$$H(Q) \leq R$$

$$\text{Choose } k^{(n)} \geq \lceil \log |A^{(n)}| \rceil$$

Then consider the encoder-decoder

pair $f^{(n)}, g^{(n)}$ s.t.

$f^{(n)}$ encodes the elements of $A^{(n)}$

to distinct strings in $\{0,1\}^{k(n)}$, and

maps strings in $X^n - A^{(n)}$ to the

string $0^{k(n)}$.

The decoding function $g^{(n)}$ is as follows.

$g^{(n)}(y^{k(n)})$ is mapped to the unique $x^n \in A^{(n)}$

satisfying $f^{(n)}(x^n) = y^{k(n)}$.

Now, note that

$$|T_Q^{(n)}| \leq 2^{nH(Q)}$$

$$\text{Then, } |A^{(n)}| = \sum_{\substack{Q \in \Omega^{(n)} : \\ H(Q) \leq R}} |T_Q^{(n)}| \leq \sum_{\substack{Q \in \Omega^{(n)} : \\ H(Q) \leq R}} 2^{nH(Q)}$$

(By Lemma)

$$\leq |\mathcal{P}^{(n)}| \cdot 2^{nR}$$

$$\leq (n+1)^{|X|} \cdot 2^{nR} \quad [\text{By Lemma}]$$

So, $\log |A^n| \leq |X| \log(n+1) + nR$

Choose $k^{(n)} = \lceil |X| \log(n+1) + nR \rceil$

Then, $|X| \frac{\log(n+1)}{n} + R \leq \frac{k^{(n)}}{n} \leq |X| \frac{\log(n+1)}{n} + R + \frac{1}{n}$

Hence $\frac{k^{(n)}}{n} \rightarrow R$.

On the other hand, note that

$$x^n \neq g^{(n)}(f^{(n)}(x^n)) \iff x^n \in X - A^n$$

$$\begin{aligned} \text{Then, } P_x(x^n \neq g^{(n)}(f^{(n)}(x^n))) \\ &= P_x(X - A^n) \\ &= P_x\left(\bigcup_{Q \in \Omega^{(n)}} T_Q^{(n)}\right) \quad [\because x^n = \bigcup_{Q \in \Omega^{(n)}} T_Q^{(n)}] \\ &\quad I(Q) > R \end{aligned}$$

$$= \sum_{P \in \mathcal{P}^{(n)}} P_x(T_Q^{(n)}) \quad (\because T_Q^{(n)} \text{ are disjoint})$$

$$H(Q) > R$$

$$\leq \sum_{Q \in \mathcal{P}^{(n)}} 2^{-n D(Q||P_x)} \quad [\text{By lemma}]$$

$$H(Q) > R$$

$$\leq 2^{-n \left[\inf_{\substack{Q \in \mathcal{P}^{(n)} : \\ H(Q) > R}} D(Q||P_x) \right]} \cdot (n+1)^{|X|}$$

$$[\because |\mathcal{P}^{(n)}|$$

$$\leq (n+1)^{|X|}]$$

$$\leq 2^{-n \left[\inf_{Q : H(Q) > R} D(Q||P_x) - |X| \log \frac{(n+1)}{n} \right]}$$

$$\approx 2^{-n C^{(n)}}$$

∴ \blacksquare

Remark:- For this theorem to be meaningful, we need $C^{(n)} > 0$ if $R > H(P_X)$. We need to build some machinery before we can prove this.

- Lemma (Data processing inequality) :-

Consider joint distributions $P_X \cdot P_{Y|X}$ and $Q_X \cdot P_{Y|X}$, and let P_Y and Q_Y be the respective marginals.

Then, for any f-divergence,

$$D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y).$$

Proof:- First, notice that

$$D_f(P_X P_{Y|X} \| Q_X P_{Y|X})$$

$$= \sum_{x \in X} \sum_{y \in Y} Q_x(x) \cdot P_{Y|X}(y|x) \cdot f\left(\frac{P_x(x) \cdot P_{Y|X}(y|x)}{Q_x(x) \cdot P_{Y|X}(y|x)}\right)$$

$$= \sum_{x \in X} Q_x(x) \cdot f\left(\frac{P_x(x)}{Q_x(x)}\right) \left(\sum_{y \in Y} P_{Y|X}(y|x) \right)$$

$$: \sum_{x \in X} Q_x(x) \cdot f\left(\frac{P_x(x)}{Q_x(x)}\right)$$

$$= D_f(P_x || Q_x) \quad \dots \dots \textcircled{1}$$

On the other hand, define

conditional distribution $P_{X|Y}, Q_{X|Y}$

such that

$$P_{X|Y} \cdot P_Y = P_x \cdot P_{Y|X}$$

$$Q_{X|Y} \cdot Q_Y = Q_x \cdot P_{Y|X}$$

Now,

$$D_f(P_x \cdot P_{Y|X} \| Q_x \cdot P_{Y|X})$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q_Y(y) \cdot Q_{X|Y}(x|y) f\left(\frac{P_Y(y) \cdot P_{X|Y}(x|y)}{Q_Y(y) \cdot Q_{X|Y}(x|y)}\right)$$

$$= \sum_{y \in \mathcal{Y}} Q_Y(y) \cdot \sum_{x \in \mathcal{X}} Q_{X|Y}(x|y) f\left(\frac{P_Y(y) \cdot P_{X|Y}(x|y)}{Q_Y(y) \cdot Q_{X|Y}(x|y)}\right)$$

Now, define the random

variables $Z_y = \frac{P_Y(y)}{Q_Y(y)} \cdot \frac{P_{X|Y}(x|y)}{Q_{X|Y}(x|y)}$

with pmf $P(Z_y = \frac{P_Y(y)}{Q_Y(y)} \cdot \frac{P_{X|Y}(x|y)}{Q_{X|Y}(x|y)})$
 $= Q_{X|Y}(x|y)$

Then, as f is convex, Jensen's inequality gives

$$\sum_{x \in \mathcal{X}} Q_{X|Y}(x|y) f(Z_y) \geq f\left(\sum_{x \in \mathcal{X}} Q_{X|Y}(x|y) \cdot Z_y\right)$$

Hence,

$$D_f(P_x \cdot P_{Y|X} || Q_x \cdot P_{Y|X})$$

$$\geq \sum_{y \in Y} Q_y(y) \cdot f\left(\sum_{x \in X} Q_{x|y}(x|y) \cdot \frac{P_y(y) \cdot P_{x|y}(x|y)}{Q_y(y) Q_{x|y}(x|y)}\right)$$

$$= \sum_{y \in Y} Q_y(y) \cdot f\left(\frac{P_y(y)}{\sum_{x \in X} Q_{x|y}(x|y)} \sum_{x \in X} P_{x|y}(x|y)\right)$$

$$= \sum_{y \in Y} Q_y(y) \cdot f\left(\frac{P_y(y)}{Q_y(y)}\right)$$

$$= D_f(P_y || Q_y) \dots \textcircled{2}$$

Thus, combining $\textcircled{1}$ and $\textcircled{2}$,
we have

$$D_f(P_x || Q_x) \geq D_f(P_y || Q_y).$$

\(\therefore \text{D}\)

Lemma :- Let $\{\Sigma_1, \dots, \Sigma_k\}$ be a partition

of X , i.e. $\bigcup_{i=1}^k \Sigma_i = X$ and $\Sigma_i \cap \Sigma_j = \emptyset$ if $i \neq j$.

Let P, Q be two fmfs on X and
define two fmfs P_Σ and Q_Σ on
 $[k] = \{1, 2, \dots, k\}$ as follows.

$$P_\Sigma(i) \triangleq P(X \in \Sigma_i)$$

$$Q_\Sigma(i) \triangleq Q(X \in \Sigma_i).$$

Then, for any f-divergence,

$$D_f(P||Q) \geq D_f(P_\Sigma || Q_\Sigma).$$

Proof :- Define $P_{I|X}(i|x) = 1$, if $x \in \Sigma_i$
 $= 0$, o.w.

Then, $P_{I|X}(\cdot|x)$ are conditional
fmfs on $[k]$.

Next note that

$$\sum_{x \in X} P_X(x) \cdot P_{I|X}(i|x)$$

$$= \sum_{x \in \Sigma_i} P_X(x)$$

$$= P_E(i)$$

Similarly $\sum_{x \in \Sigma} Q_X(x) \cdot P_{I|X}(i|x)$

$$= Q_E(i).$$

Thus, by data processing inequality

$$D_f(P||Q) \geq D_f(P_E || Q_E)$$

..... A

• Lemma (Pinaki inequality) :-

$$D(P||Q) \geq \frac{2}{\ln 2} TV(P, Q)^2.$$

Proof :- We first prove this for $|X|=2$.

For $|X|=2$, let $P = B_2(p)$ and
 $Q = B_2(q)$. Then,

$$\begin{aligned} D(P||Q) &= \frac{2}{\ln 2} TV^2(P, Q) \\ &= p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{(1-q)} \\ &\quad - \frac{1}{2 \ln 2} ((|p-q| + |1-p-q|)^2 \\ &= p \log \frac{p}{q} + (1-p) \cdot \log \frac{(1-p)}{(1-q)} - \frac{2}{\ln 2} (p-q)^2 \end{aligned}$$

Now, define

$$\begin{aligned} g(p, q) &= p \log \frac{p}{q} + (1-p) \log \frac{(1-p)}{1-q} \\ &\quad - \frac{2}{\ln 2} (p-q)^2. \end{aligned}$$

Note that if $p=q$, $g(p, q) = 0$.

Now, observe that

$$\begin{aligned}\frac{\partial g(p, q)}{\partial q} &= -\frac{p}{q \ln 2} \cdot \frac{p}{p} \cdot qr \\ &\quad + \frac{(1-p)}{(1-q)\ln 2} \cdot \frac{(1-p)}{(1-p)} \cdot (1-q) \\ &\quad + \frac{4}{\ln 2} (p-q) \\ &= \frac{1}{\ln 2} \left[\frac{1-p}{1-q} - \frac{p}{q} + 4(p-q) \right] \\ &= \frac{1}{(1-q) \cdot q \cdot \ln 2} \left[(q-p) + 4q(1-q)(p-q) \right] \\ &= \frac{(p-q)}{(1-q)q \ln 2} \left[4q(1-q) - 1 \right]\end{aligned}$$

Noting that $4q(1-q) \leq 1$,

when $f \geq g$, we have

$$\frac{\partial g(f, g)}{\partial g} \leq 0$$

Then, since $g(f, f) = 0$, we have

$$g(f, g) \geq 0, \text{ if } g \leq f.$$

Alternatively, assume $f \leq g$.

Then, $\frac{\partial g(f, g)}{\partial g} \geq 0$

Then, since $g(f, f) = 0$, we have

$$g(f, g) \geq 0, \text{ if } f \leq g.$$

Hence, $g(f, g) \geq 0$, for f, g and

thus, if $|D| = 2$,

$$D(P||Q) \geq \frac{2}{\ln 2} TV^2(P, Q).$$

Now, for any X , define $A \subseteq X$ as

$$A = \{x : P(x) \geq Q(x)\}$$

Then $\{E_1, E_2\}$ is a partition of X ,

where $E_1 = A$, $E_2 = A^c$.

Thus, using the Pinsker inequality for binary random variables and the previous lemma, we have

$$D(P||Q) \geq D(P_E || Q_E)$$

$$\geq \frac{2}{\ln 2} TV^2(P_E || Q_E)$$

Now, $TV(P_E || Q_E)$

$$= \frac{1}{2} \left[|P(A) - Q(A)| + |P(A^c) - Q(A^c)| \right]$$

$$= \frac{1}{2} \left[\left| \sum_{x \in A} (P(x) - Q(x)) \right| + \left| \sum_{x \in A^c} (P(x) - Q(x)) \right| \right]$$

$$= \frac{1}{2} \left[\sum_{x \in A} |P(x) - Q(x)| + \sum_{x \in A^c} |P(x) - Q(x)| \right]$$

[By definition A,

$$P(x) \geq Q(x) \quad \forall x \in A$$

$$\text{and } Q(x) > P(x) \quad \forall x \in A^c,$$

and hence the Δ
inequalities are satisfied
with equality]

$$= \frac{1}{2} \sum_{x \in X} |P(x) - Q(x)|$$

$$= TV(P, Q).$$

$$\text{Thus, } D(P||Q) \geq \frac{2}{\ln 2} TV^2(P, Q)$$

..... 13

Lemma :- Consider the function

$$f(t) = -t \log t, \quad \forall t \in [0, 1]. \text{ Then,}$$

(i) f is concave.

(ii) For any $0 \leq \tau \leq \frac{1}{2}$ and any

$$0 \leq t \leq 1-\tau,$$

$$\begin{aligned} |f(t+\tau) - f(t)| &\leq \max \{f(\tau), f(1-\tau)\} \\ &= -\tau \log \tau. \end{aligned}$$

Proof :- (i) $f'(t) = -\log e - \log t$

$$f''(t) = -\frac{1}{t} \log e$$

Thus, $f''(t) < 0$ when $t > 0$

and $f(0) = \lim_{t \rightarrow 0^+} f(t)$.

Thus, f is concave.

(ii) We shall assume $\tau > 0$, as otherwise it is trivial.

Consider any $[x_1, x_2] \subseteq [0, 1]$ and

let $x_1 < x < x_2$.

Then, similar to the proof where we established that a function is convex iff its derivative is non-decreasing, we can show using the concavity of f that

$$\frac{f(x_2) - f(x)}{x_2 - x} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{f(x) - f(x_1)}{x - x_1}$$

... ①

Now, consider any four points

$$x_1 \leq x_2 \leq x_3 \leq x_4.$$

We claim that

$$\frac{f(x_3) - f(x_1)}{x_3 - x_1} \geq \frac{f(x_4) - f(x_2)}{x_4 - x_2}$$

... ②

To see why ② holds,

Take $x_1' = x_1$, $x' = x_2$, $x_2' = x_3$ and

thus $x' \in [x_1', x_2']$.

Then by ①, we have

$$\frac{f(x_3) - f(x_1)}{x_3 - x_1} \geq \frac{f(x_2) - f(x_1)}{x_3 - x_1} \dots @$$

On the other hand, choosing

$x_1' = x_2$, $x' = x_3$, and $x_2' = x_4$,

we have $x' \in [x_1', x_2']$, and

hence by ①, we have

$$\frac{f(x_3) - f(x_2)}{x_3 - x_2} \geq \frac{f(x_4) - f(x_2)}{x_4 - x_2} \dots ①.$$

Combining ① and ② we have

$$\frac{f(x_3) - f(x_1)}{x_3 - x_1} \geq \frac{f(x_4) - f(x_2)}{x_4 - x_2}, \text{ which}$$

proves ②.

Next, we claim the following:

$$\frac{f(x_4) - f(x_3)}{x_4 - x_3} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \dots \textcircled{3}$$

Again to prove this, we first take

$x'_1 = x_1$, $x' = x_2$, $x'_2 = x_3$ and hence by ①

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq \frac{f(x_3) - f(x_2)}{x_3 - x_2}$$

Similarly, taking $x'_1 = x_2$, $x' = x_3$ and $x'_2 = x_4$,

we get using ① that

$$\frac{f(x_3) - f(x_2)}{x_3 - x_2} \geq \frac{f(x_4) - f(x_3)}{x_4 - x_3}.$$

Combining, we get

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq \frac{f(x_4) - f(x_3)}{x_4 - x_3}$$

which proves ③.

Now, note that if $\tilde{\tau} \leq \frac{1}{2}$, then

$$1 - \tilde{\tau} \geq \tau.$$

For any $0 \leq t \leq 1 - \tilde{\tau}$, define

$$x_1 = 0, x_2 = \min\{t, \tau\}, x_3 = \max\{t, \tau\}.$$

$$x_4 = t + \tilde{\tau}.$$

Then, we have by ② and ③

$$\frac{f(t + \tilde{\tau}) - f(t)}{\tilde{\tau}} \leq \frac{f(\tau) - f(0)}{\tau} = \frac{f(\tau)}{\tau} \quad [\because f'(0) = 0]$$

$$\text{Thus, } f(t + \tilde{\tau}) - f(t) \leq f(\tau)$$

$$\leq \max\{f(\tau), f(1 - \tilde{\tau})\}$$

..... ④

On the other hand, choose

$$x_1 = t, \quad x_2 = \min \{t+\tau, 1-\tau\}, \quad x_3 = \max \{t+\tau, 1-\tau\},$$

$x_4 = 1$. Then by ② and ③,

we have

$$\begin{aligned} \frac{f(t+\tau) - f(t)}{\tau} &\geq \frac{f(1) - f(1-\tau)}{\tau} \\ &= -\frac{f'(1-\tau)}{\tau} \quad (\because f'(1)=0) \end{aligned}$$

Thus, we have

$$\begin{aligned} f(t+\tau) - f(t) &\geq -f'(1-\tau) \\ &\geq -\max \{f(\tau), f(1-\tau)\} \end{aligned}$$

Combining ④ and ⑤, we have ... ⑤

$$|f(t+\tau) - f(t)| \leq \max \{f(\tau), f(1-\tau)\}$$

It remains to prove

$$f(\tau) \geq f(1-\tau) \quad \text{for } 0 \leq \tau \leq \frac{1}{2}.$$

To see this, define $g(\tau) = f(\tau) - f(1-\tau)$

Then, $g(0) = f(0) - f(1) = 0$.

$$\begin{aligned} g'(\tau) &= f'(\tau) + f'(1-\tau) \\ &= -2\log e - \log(\tau \cdot (1-\tau)) \end{aligned}$$

$$g''(\tau) = -\frac{\log e}{\tau(1-\tau)}(1-2\tau)$$

$$\leq 0, \quad \text{when } 0 \leq \tau \leq \frac{1}{2}$$

Thus, $g(\tau)$ is concave in τ .

Also, $g(0) = g\left(\frac{1}{2}\right) = 0$.

Thus, by ①, we have

$$\frac{g(\tau) - 0}{\tau} \geq \frac{g\left(\frac{1}{2}\right) - g(0)}{\frac{1}{2}} = 0.$$

$\Rightarrow g(\tau) \geq 0$, and hence $f(\tau) \geq f(1-\tau) \dots \square$

Lemma (Continuity of entropy): -

Let $TV(P, Q) \leq \frac{1}{4}$. Then

$$|H(P) - H(Q)| \leq -2TV(P, Q) \log\left(\frac{2TV(P, Q)}{|X|}\right).$$

Proof:- $|H(P) - H(Q)|$

$$= \left| \sum_{x \in X} P(x) \log P(x) - \sum_{x \in X} Q(x) \log Q(x) \right|$$

$$\leq \sum_{x \in X} |P(x) \log P(x) - Q(x) \log Q(x)|$$

$$= \sum_{x \in X} |\beta(P(x)) - \beta(Q(x))| \dots \textcircled{1}$$

where $\beta(t) = -t \log t$.

$$\text{Now, } TV(P, Q) = \frac{1}{2} \sum_{x \in X} |P(x) - Q(x)|$$

$$\text{Then } TV(P, Q) \leq \frac{1}{4}$$

$$\Rightarrow \frac{1}{2} \sum_{x \in \Sigma} |P(x) - Q(x)| \leq \frac{1}{4}$$

$$\Rightarrow \sum_{x \in \Sigma} |P(x) - Q(x)| \leq \frac{1}{2}$$

$\therefore |P(x) - Q(x)| \geq 0, \forall x \in \Sigma$, we

must have

$$|P(x) - Q(x)| \leq \frac{1}{2}, \forall x \in \Sigma.$$

In the previous lemma,

$$\text{use } t + \bar{t} = \max \{P(x), Q(x)\}$$

$$\text{and } t = \min \{P(x), Q(x)\}$$

Note that as $|P(x) - Q(x)| \leq \frac{1}{2}$,

$$\bar{t} \leq \frac{1}{2}.$$

$$\text{Also, } 1 - |P(x) - Q(x)|$$

$$= 1 - \max \{P(x), Q(x)\} + \min \{P(x), Q(x)\}$$

$$\geq \min \{P(x), Q(x)\},$$

i.e., $|-\tau| \geq t$.

Thus, by the previous lemma,
we have

$$|f(P(x)) - f(Q(x))| \leq -|P(x) - Q(x)| \log |P(x) - Q(x)|$$

$\forall x \in X.$

Plugging this in ①, we get

$$|H(P) - H(Q)|$$

$$\leq - \sum_{x \in X} |P(x) - Q(x)| \log |P(x) - Q(x)|$$

$$= -2TV(P, Q) \left[\sum_{x \in X} \frac{|P(x) - Q(x)|}{2TV(P, Q)} \log \frac{|P(x) - Q(x)|}{2TV(P, Q)} \right]$$

$$+ \sum_{x \in X} \frac{|P(x) - Q(x)|}{2TV(P, Q)} \log (2TV(P, Q)) \Big]$$

$$\leq 2TV(P, Q) \log |X| - 2TV(P, Q) \log (2TV(P, Q))$$

$\therefore \frac{|P(x) - Q(x)|}{2TV(P, Q)}$ is a fmf on X

and $H\left(\frac{|P(\cdot) - Q(\cdot)|}{2TV(P, Q)}\right) \leq \log |X|$

$$= 2TV(P, Q) \log\left(\frac{|X|}{2TV(P, Q)}\right)$$

..... \square

Theorem :- The universal source coding

exponent $C^{(n)} \geq_n 0$ if $H(P_X) < R$.

Proof :- We shall prove that

$$\inf_{\substack{Q: H(Q) > R}} D(Q||P_X) > 0 \text{ if } R > H(P_X).$$

Then, noting that

$$C^{(n)} = \inf_{\substack{Q: H(Q) > R}} D(Q||P_X) - |X| \log \frac{(n+1)}{n},$$

we have $C^{(n)} \geq_n 0$ since

$$\frac{\log(n+1)}{n} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

So, assume that

$$\inf_{Q: H(Q) > R} D(Q || P_X) = 0 \quad \text{if } R > H(P_X)$$

Now, this implies that

\exists a sequence $Q_n, n \geq 1$, of pmfs on

X satisfying $D(Q_n || P_X) \rightarrow 0$,

and $H(Q_n) > R \quad \forall n$.

Now, by Pinsker's inequality,

$$D(Q_n || P_X) \geq \frac{2}{\ln 2} TV(Q_n, P_X)^2 \geq 0$$

Thus, by a sandwiching argument,

we have $TV(Q_n, P_X) \rightarrow 0$.

$$\text{Now, } \varepsilon = \frac{R - H(P_x)}{2} > 0.$$

Continuity of entropy tells us that

$$|H(Q_n) - H(P_x)| \leq_n 2TV(Q_n, P_x) \log \frac{|x|}{2TV(Q_n, P_x)}$$

$$\text{as } TV(Q_n, P_x) \leq_n \frac{1}{4}.$$

$$\text{Next, note that } f(x) = x \log \frac{|x|}{x}$$

is continuous at $x=0$, if we

$$\text{define } f(0) = 0.$$

$$\text{Hence, } \exists \delta > 0, \text{ s.t. } |f(x) - f(0)| \leq \frac{\varepsilon}{2}$$

$$\text{whenever } |x| \leq \delta$$

$$\text{Then, noting that } TV(Q_n, P_x) \rightarrow 0,$$

$$\text{we have } TV(Q_n, P_x) \leq_n \delta.$$

$$\text{Hence, } 2TV(Q_n, P_x) \log \frac{|x|}{2TV(Q_n, P_x)} \leq_n \frac{\varepsilon}{2}.$$

$$\Rightarrow |H(Q_n) - H(P_x)| \leq \frac{\varepsilon}{2}.$$

$$\Rightarrow H(Q_n) \leq_n H(P_x) + \frac{\varepsilon}{2}$$

$$< R$$

This violates the fact that

$H(Q_n) > R$ and hence our assumption that

$\inf_{Q: H(Q) > R} D(Q||P_x) = 0$ if $R > H(P_x)$ was wrong.

Thus, by non-negativity of relative entropy, we must have

$\inf_{Q: H(Q) > R} D(Q||P_x) > 0$ if $R > H(P_x)$ and

hence, $C_n \geq 0$ if $R > H(P_x)$.

... B

* Information Theory & counting

- Lemma (Han's inequality):

$$H(X^n) \leq \frac{1}{(n-1)} \sum_{i=1}^n H(X_{\neq i}),$$

where $X_{\neq i} \triangleq X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$.

Proof :-

$$\begin{aligned} H(X^n) &= H(X_{\neq i}) + H(X_i | X_{\neq i}) \\ &\leq H(X_{\neq i}) + H(X_i | X^{i-1}) \end{aligned}$$

[\because Conditioning reduces entropy]

Then,

$$\begin{aligned} nH(X^n) &\leq \sum_{i=1}^n H(X_{\neq i}) + \sum_{i=1}^n H(X_i | X^{i-1}) \\ &= \sum_{i=1}^n H(X_{\neq i}) + H(X^n) \end{aligned}$$

$$\Rightarrow H(X^n) \leq \frac{1}{(n-1)} \sum_{i=1}^n H(X_{\neq i}) \quad \dots B$$

Definition :- Let $x, y \in \{0,1\}^n$. Then, the Hamming distance between x and y is given by

$$d_H(x, y) = |\{i \in [n] : x_i \neq y_i\}|.$$

Definition :- A binary hypercube is a graph $G = (V, E)$, where $V \subseteq \{0,1\}^n$, and $(x, y) \in E$ if $d_H(x, y) = 1$.

[Note:- For those interested, a binary hypercube is a Cayley graph with generating set $S = \{e_i : i \in [n]\}$, where e_i is the n -bit string with one only in the i th position. The Cayley graph with any $S \subseteq \{0,1\}^n$ is called a generalised hypercube.]

• Definition:- Let $A \subseteq \{0,1\}^n$. Then,
 the boundary of A in the hypercube
 is given by $\partial A \triangleq \{x \in A^c : \exists y \in A$
 s.t. $(x,y) \in E\}$.

• Theorem (Edge isoperimetric inequality
 for the hypercube):
 $|\partial A| \geq n|A| - |A|\log|A|$.

Proof:- Define $E_A \triangleq \{(x,y) \in E :
 x \in A, y \in A\}$.

Now, let X^n be the
 co-ordinates of a vertex chosen
 uniformly at random from A .

Then, $P(X^n = x) = \frac{1\{x \in A\}}{|A|}$.

Now, for any $x \in A$, and any $i \in [n]$

$$\begin{aligned} P(x_i | x_{\neq i}) &= 1, \text{ if } x^{\bar{i}} \notin A \\ &= \frac{1}{a}, \text{ if } x^{\bar{i}} \in A, \end{aligned}$$

where $x^{\bar{i}} \triangleq x_1, \dots, x_{i-1}, \bar{x}_i, x_{i+1}, \dots, x_n$.

Now, $H(X_i | X_{\neq i})$

$$= \sum_{x \in A} P(x) \log \frac{1}{P(x_i | x_{\neq i})}$$

$$= \sum_{x \in A} P(x) \quad 1\{x^{\bar{i}} \in A\}$$

$$= \frac{1}{|A|} \sum_{x \in A} 1\{x^{\bar{i}} \in A\} \quad \dots \textcircled{1}$$

Now, note that $(x, y) \in E$ only if $\exists i \in \{n\}$

s.t. $y = x^i$.

$$\text{Thus, } |E_A| = \frac{1}{2} \sum_{x \in A} \sum_{i=1}^n 1\{|x^i \in A\} \dots \textcircled{2}$$

So, from \textcircled{1} & \textcircled{2}

$$\sum_{i=1}^n H(x_i | X_{\neq i}) = \frac{2}{|A|} \cdot |E_A|$$

Now, by Hans inequality, we have

$$n H(X^n) - \sum_{i=1}^n H(X_{\neq i}) \leq H(X^n)$$

$$\Rightarrow \sum_{i=1}^n H(x_i | X_{\neq i}) \leq H(X^n)$$

$$\Rightarrow \frac{2}{|A|} |E_A| \leq \log |A|$$

$$\Rightarrow |E_A| \geq \frac{|A| \log |A|}{2}.$$

Now,

$$|\partial A| = \sum_{x \in A} \sum_{i=1}^n 1\{x^i \notin A\}$$

$$= \sum_{x \in A} \sum_{i=1}^n (1 - 1\{x^i \in A\})$$

$$= n|A| - 2|E_A|$$

$$\geq n|A| - |A| \log |A|$$

..... β

Theorem (Shearer inequality- counting form) :- Let S_1, \dots, S_k

be a collection of subsets of $[n]$

s.t. every $i \in [n]$ is contained in at least t of these sets S_1, \dots, S_k .

Then, for any random variables

$$X_1, \dots, X_n,$$

$$tH(X^n) \leq \sum_{i=1}^k H(X_{S_k}).$$

Proof :-

$$\sum_{i=1}^k H(X_{S_k})$$

$$= \sum_{i=1}^k \sum_{j \in S_i} H(X_j | X_{\{l \in S_k : l < j\}})$$

$$\geq \sum_{i=1}^k \sum_{j \in S_i} H(X_j | X^{j-1})$$

[Conditioning

$$= \sum_{i=1}^k \sum_{j=1}^n H(X_j | X^{j-1}) \cdot 1_{\{j \in S_i\}}$$

reduces entropy]

$$= \sum_{j=1}^n \sum_{i=1}^k H(X_j | X^{j-1}) \cdot 1_{\{j \in S_i\}}$$

$$\geq \sum_{j=1}^n t \cdot H(X_j | X^{j-1})$$

(By definition
of S_1, \dots, S_k)

$$= t H(X^n) \dots \blacksquare$$

Theorem ('Discrete Loomis Whitney Inequality'): Let S be a finite collection of points in \mathbb{R}^d .

Define projection maps

$f_{\pm i}: \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$, by

$f_{\pm i}(x) = x_{\pm i}$.

Let $f_{\pm i}(S) \triangleq \{y \in \mathbb{R}^{n-1} : \exists x \in S \text{ s.t. } f_{\pm i}(x) = y\}$.

Suppose $|f_{\pm i}(S)| = n_i$.

Then, $|S| \leq \prod_{i=1}^d n_i^{\frac{1}{d-1}}$

Proof:- Let X^d be the co-ordinate of a point in S chosen uniformly at random.

$$\text{Then, } H(X^d) = \log |S|$$

Now, note that $X_{\neq i}$ takes values in $\mathbb{f}_{\neq i}(S)$.

$$\text{Thus, } H(X_{\neq i}) \leq \log n_i .$$

Now, consider the collection of subsets of $[d]$ given by T_1, \dots, T_d where

$$T_i = [d] \setminus \{i\} .$$

Then every $i \in [d]$ appears in exactly $d-1$ sets of T_1, \dots, T_d .

Then, by Shearer's inequality/Han's inequality,

$$(d-1) \cdot H(x^d) \leq \sum_{i=1}^d H(X_{\pm i})$$
$$= \sum_{i=1}^d H(X_{\pm i})$$

$$\Rightarrow (d-1) \cdot \log |S| \leq \sum_{i=1}^d \log n_i$$

$$\Rightarrow \log |S|^{d-1} \leq \log \left(\prod_{i=1}^d n_i \right)$$

$$\Rightarrow |S| \leq \prod_{i=1}^d n_i^{\frac{1}{d-1}}$$

..... \square

Theorem (Shearer Inequality Probabilistic form): Let P be a distribution on $2^{[n]}$ s.t. for any $i \in [n]$,

$P(i \in S) \geq \mu$. Then, for any jointly distributed random variables

$$X^n, \quad \mu H(X^n) \leq E_p[H(X_s)].$$

Proof :- $E_p[H(X_s)]$

$$= E_p\left[\sum_{i \in s} H(X_i | X_{\{j : j < i\}})\right]$$

$$\geq E_p\left[\sum_{i \in s} H(X_i | X^{i-1})\right]$$

[\because Conditioning
reduces entropy &
Expectation is order
preserving]

$$= E_p\left[\sum_{i=1}^n H(X_i | X^{i-1}) \cdot 1_{\{i \in s\}}\right]$$

$$= \sum_{i=1}^n H(X_i | X^{i-1}) E_p[1_{\{i \in s\}}]$$

$$= \sum_{i=1}^n H(X_i | X^{i-1}) \cdot P(i \in s)$$

$$\geq \mu \cdot \sum_{i=1}^n H(X_i | X^{i-1})$$

$$= \mu H(X^n) \quad \dots \quad \square$$

Remark:- We obtain the counting version of the Shearer lemma from this version as follows. Choose P.s.t.

$$P(S_j) = \frac{1}{k}, \quad \forall j \in [k]$$

and $P(S) = 0$, if $S \in \{S_1, \dots, S_k\}$

$$\text{Then, } P(i \in S) = \frac{1}{k} \cdot \sum_{j=1}^k \mathbb{1}_{\{i \in S_j\}} \geq \frac{t}{k}.$$

$$\text{Also, } \mathbb{E}_P[H(X_S)]$$

$$= \frac{1}{k} \sum_{j=1}^k H(S_j)$$

Thus, using the probabilistic form of Shearer's lemma, $\frac{1}{k} \sum_{j=1}^k H(S_j) \geq \frac{t}{k} \cdot H(X^n)$.

• Definition :- For a graph $G = (V, E)$,
the set of vertices $\{x_1, x_2, x_3\}$ form a triangle
if $(x_1, x_2), (x_2, x_3), (x_3, x_1) \in E$.

• Theorem (Triangles versus edges):-

For a graph $G = (V, E)$, let

T denote the set of all its triangles.

Then, $|E| \geq \frac{1}{2} (6|T|)^{2/3}$.

Proof: Define the random
variables X_1, X_2, X_3 as follow

$$P(X_1, X_2, X_3 = x_1, x_2, x_3) = \frac{1}{6|T|},$$

if $\{x_1, x_2, x_3\} \in T$

$= 0, \text{ e.w.}$

Note that this is well-defined since

$$\sum_{(x_1, x_2, x_3) \in V^3} P(x_1, x_2, x_3 = x_1, x_2, x_3)$$

$$= \sum_{\substack{\{t_1, t_2, t_3\} \\ \in T}} 6 \cdot P(x_1, x_2, x_3 = t_1, t_2, t_3)$$

[$\because X_1, X_2, X_3$ taking values in all 6 permutations of (t_1, t_2, t_3) is included in the original sum]

$$= 1.$$

Now, let Q be a distribution

on $2^{[3]}$ defined by

$$Q(S) = 0, |S| \neq 2$$

$$= \frac{1}{3}, |S|=2.$$

Again, note that this is well defined since

$$|\{S \in 2^{\{3\}} : |S|=2\}| = \binom{3}{2} = 3$$

Thus, $Q(i \in S) = \frac{2}{3}$, $\forall i \in \{1, 2, 3\}$

Then, by Shearer inequality, we have,

$$\frac{2}{3} H(X_1, X_2, X_3) \leq \mathbb{E}_Q[H(X_S)]$$

Since $Q(S) \geq 0$, only if $|S|=2$,

$$\text{and } \mathbb{E}_Q[H(X_S)] \geq \frac{2}{3} H(X_1, X_2, X_3),$$

$\exists S^* \subseteq \{1, 2, 3\}$, s.t. $|S^*|=2$ satisfying

$$\begin{aligned} H(X_{S^*}) &\geq \frac{2}{3} H(X_1, X_2, X_3) \\ &= \frac{2}{3} \cdot \log(6/\pi) \dots \textcircled{1} \end{aligned}$$

Now, since X_1, X_2, X_3 puts mass

(i.e. have non-zero probability)

only on tuples (x_1, x_2, x_3) s.t.

$$\{x_1, x_2, x_3\} \in T.$$

Then for any S with $|S|=2$,

if X_S puts mass on the tuple

$$(x_1, x_2) \text{ then } \{x_1, x_2\} \in E$$

Thus, $|\text{support}(X_S)| \leq 2|E|$

[where 2 takes
care of the permutations
 $(x_1, x_2) \leftrightarrow (x_2, x_1)$]

Thus $H(X_{S^*}) \leq \log |\text{support}(X_S)|$

$$\leq \log(2|E|)$$

Plugging this in ①, we have

$$\log(2|E|) \geq \frac{2}{3} \log(6|T|)$$
$$\Rightarrow |E| \geq \frac{1}{2} (6|T|)^{\frac{2}{3}}$$

... 

* Mutual information

Let X, Y have a joint pmf P_{XY} , with P_X and P_Y being corresponding marginals.

Mutual information

$$I(X;Y) \triangleq D(P_{XY} \| P_X P_Y)$$

In other words

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)}$$
$$= E_{P_{XY}} \left[\log \frac{P_{XY}(X,Y)}{P_X(X) \cdot P_Y(Y)} \right]$$

Remark:-

An intuitive understanding of mutual information is that it is the amount of information common to X and Y .

However, as we have seen several times in this course, such an understanding, although intuition providing, is vague and immature.

In fact, there exist other measures of 'information common to X and Y '

such as Wyner common information
and Gracs-Körner common information.

Properties of mutual information

• Lemma :-

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

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

$$iii) I(X; Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

$$iv) I(X; X) = H(X)$$

$$v) I(X; Y) = 0 \text{ iff } X \perp Y.$$

$$vi) I(f(X); X) = H(f(X))$$

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

Proof! :- i), ii), vi) follow immediately
from the defn and the
nm-negativity of relative entropy.

iii), vii)

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)} \\ &= \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{X|Y}(x|y)}{P_X(x)} \\ &= \sum_x \sum_y P_{XY}(x,y) \log P_{X|Y}(x|y) \\ &\quad + \sum_x \left(\sum_y P_{XY}(x,y) \right) \log \frac{1}{P_X(x)} \\ &= -H(X|Y) + \sum_x P_X(x) \log \frac{1}{P_X(x)} \\ &= -H(X|Y) + H(X) \end{aligned}$$

Similarly, we can show

$$I(X;Y) = H(Y) - H(Y|X).$$

Then, by chain rule, we have

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

$$\text{iv)} I(x; x) = H(x) - H(x|x) \quad (\text{by part (iii)}) \\ = H(x)$$

$$[\because \text{Given } X=x, P_{x|x=x}(x'=x) \\ = 1\{x'=x\}.]$$

$$\text{v)} I(f(x); X) = H(f(x)) - H(f(x)|x) \quad (\text{By part (iii)}) \\ = H(f(x)) \quad [\because H(f(x)|X) = 0]$$

..... ☐

* Conditional relative entropy,
mutual information, and chain
rule

Let $P_{Y|X}, Q_{Y|X}$ be two conditional
pmf on Y , and let P_X be a
pmf on X .

$$\text{Then, } D(P_{Y|X} || Q_{Y|X} | P_X)$$

$$= \mathbb{E}_{P_X} \left[\sum_y P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \right]$$

• Lemma (chain rule): - Let P_{XY}, Q_{XY} be two joint distributions with $P_{Y|X}, Q_{Y|X}$ being the respective conditional pmfs and P_X, Q_X being the corresponding X-marginals. Then,

$$D(P_{XY} || Q_{XY}) = D(P_X || Q_X) + \\ D(P_{X|Y} || Q_{X|Y} | P_X)$$

Proof: - $D(P_{XY} || Q_{XY}) = \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$

$$\begin{aligned}
&= \sum_x \sum_y P_x(x) \cdot P_{Y|X}(y|x) \log \frac{P_x(x) \cdot P_{Y|X}(y|x)}{Q_x(x) \cdot Q_{Y|X}(y|x)} \\
&= \sum_x \sum_y P_x(x) \cdot P_{Y|X}(y|x) \log \frac{P_x(x)}{Q_x(x)} \\
&\quad + \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \\
&= \sum_x P_x(x) \cdot \log \frac{P_x(x)}{Q_x(x)} \cdot \left(\sum_y P_{Y|X}(y|x) \right) \\
&\quad + D(P_{X|Y} || Q_{X|Y} | P_x) \\
&= D(P_x || Q_x) + D(P_{X|Y} || Q_{X|Y} | P_x).
\end{aligned}$$

.. B

Let X, Y, Z be three random variables.
with joint distribution P_{XYZ}
(And corresponding marginals and
conditionals denoted accordingly).

Then,

$$I(X;Y|Z) \triangleq E_{P_Z} [I(X;Y|Z=z)]$$

$$= \sum_z P_Z(z) \cdot I(X;Y|Z=z)$$

$$= \sum_z P_Z(z) \cdot \sum_{x,y} P_{X,Y|Z}(x,y|z) \cdot$$

$$\log \left(\frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z) \cdot P_{Y|Z}(y|z)} \right)$$

- **Lemma :-**
 - i) $I(X;Y|Z) = I(Y;X|Z)$
 - ii) $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
 - iii) $I(X;Y|Z) = 0 \text{ iff } X \perp Y|Z$

Proof :- i), iii) Follows from definition
and symmetry of $I(X;Y)$.
And $I(X;Y) = 0 \Leftrightarrow X \perp Y$.

$$\begin{aligned}
 \text{ii)} I(X;Y|Z) &= \sum_z P_z(z) \cdot \sum_{x,y} P_{XYZ}(x,y|z) \cdot \\
 &\quad \log \frac{P_{XYZ}(x,y|z)}{P_{X|Z}(x|z) \cdot P_{Y|Z}(y|z)} \\
 &= \sum_z \sum_x \sum_y P_z(z) \cdot P_{XYZ}(x,y|z) \\
 &\quad \log \frac{P_{X|YZ}(x|y,z)}{P_{X|Z}(x|z)} \\
 &= \sum_z \sum_x P_z(z) P_{X|Z}(x|z) \log \frac{1}{P_{X|Z}(x|z)} \cdot \\
 &\quad \left(\sum_y P_{Y|XZ}(y|x,z) \right) \\
 &\quad - \sum_z \sum_y P_z(z) \cdot P_{Y|Z}(y|z) \\
 &\quad \sum_x P_{X|YZ}(x|y,z) \log \frac{1}{P_{X|YZ}(x|y,z)} \\
 &= H(X|Z) - H(X|Y,Z)
 \end{aligned}$$

The other equality follows
similarly. Θ

Remark:- $I(x; \gamma/z)$ and $I(x; \gamma)$ are not necessarily greater or lesser than the other.

Example 1:- $I(x; \gamma/z) > I(x; \gamma)$

Let $X, Y \sim \text{i.i.d } B_e(\frac{1}{2})$

$$Z = X \oplus Y$$

Then, $I(x; \gamma) = 0$

But, given $Z=z$,

$$P_{X|Z}(x|z) = \frac{1}{2}, \quad \forall x \in \{0, 1\}, \quad z \in \{0, 1\}$$

$$P_{X|YZ}(x|y, z) = 1\{x \oplus y = z\},$$

$$\forall y \in \{0, 1\}, \quad z \in \{0, 1\}$$

Thus, $H(X|z) = 1$

$$H(X|Y, z) = 0, \quad \text{and so}$$

$$I(x; \gamma/z) = 1.$$

Example 2:- $I(X;Y) > I(X;Y|Z)$

Let $Z \sim Be(\mu)$. Let X, Y be binary valued s.t.

$X \perp Y | Z$, and $X|Z=0, Y|Z=0 \sim Be(\mu)$,
 $X|Z=1, Y|Z=1 \sim Be(\nu)$

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

Now, $P_{XY}(x,y)$

$$\begin{aligned} &= \mu P_{XY|Z}(x,y|1) + (1-\mu) \cdot P_{XY|Z}(x,y|0) \\ &= \mu \cdot P_{X|Z}(x|1) \cdot P_{Y|Z}(y|1) \\ &\quad + (1-\mu) \cdot P_{X|Z}(x|0) \cdot P_{Y|Z}(y|0) \end{aligned}$$

$$\text{So, } P_{XY}(0,0) = \mu (1-\mu)^2 + (1-\mu)(1-\nu)^2$$

$$\begin{aligned} P_{XY}(0,1) &= \mu \mu (1-\nu) + (1-\mu) \cdot \nu (1-\mu) \\ &= P_{XY}(1,0) \end{aligned}$$

$$P_{XY}(1,1) = \mu \nu^2 + (1-\mu) \cdot \nu^2$$

Thus,

$$P_X(0) = p(1-n) + (1-p)(1-q) = P_Y(0)$$

$$P_X(1) = pn + (1-p)q = P_Y(1)$$

Take $p = \frac{1}{2}$, $n = \frac{1}{3}$, $q = \frac{1}{4}$

$$H(X) = h\left(\frac{3}{8}\right) = 0.954$$

$$\text{Also, } P_{X|Y}(0|0) = \frac{\frac{1}{4} + \frac{9}{16}}{\frac{1}{2} + \frac{3}{4}} = \frac{13}{20}$$

$$P_{X|Y}(1|0) = \frac{7}{20}$$

$$P_{X|Y}(0|1) = \frac{\frac{1}{4} + \frac{3}{16}}{\frac{1}{2} + \frac{1}{4}} = \frac{7}{12}$$

$$P_{X|Y}(1|1) = \frac{5}{12}$$

Thus, $H(X|Y) = P_X(0) \cdot h\left(\frac{7}{20}\right) + P_Y(1)h\left(\frac{5}{12}\right)$

$$\begin{aligned} &= \frac{5}{8}h\left(\frac{7}{20}\right) + \frac{3}{8}h\left(\frac{5}{12}\right) \\ &= 0.951 \end{aligned}$$

Thus, $I(X;Y) > 0$.

• Lemma $\vdash |I(X;Y) - I(X;Y|Z)| \leq H(Z)$

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

$$= H(X) - H(X|Y) - H(X|Z) - H(Z) \\ + H(X|Y, Z)$$

$$= H(X) - H(X|Y) - H(X,Z) \\ + H(X|Y, Z)$$

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

$$\leq -H(Z|X) \quad (\because \text{Conditioning reduces entropy}) \\ \leq 0$$

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

$$= H(X) - H(X|Y) - H(X|Z) \\ + H(X|Y, Z) + H(Z)$$

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

$$\geq I(X;Z) + H(Z) - H(Z|Y) \quad [\because I(X;Z|Y) \leq H(Z|Y)]$$

$$= I(X;Z) + I(Y;Z)$$

≥ 0

. 

• Lemma (Chain rule):-

$$I(X;Y^n) = \sum_{i=1}^n I(X;Y_i | Y^{i-1})$$

Proof:- $I(X;Y^n)$

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

$$= \sum_{i=1}^n [H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X)]$$

(Chain rule of entropy)

$$= \sum_{i=1}^n I(X;Y_i | Y^{i-1})$$

. 

Lemma:- Let X be a linear space and $A \subseteq X$ be a convex set. Let $f: A \rightarrow \mathbb{R}$ be a convex (resp. concave) map and $g: A \rightarrow \mathbb{R}$ be linear. Then $f+g$ is convex (resp. concave).

Proof:- We shall show the case where f is convex. The case with concave f follows similarly.

Let $x, y \in A$ and pick any $\lambda \in [0, 1]$.

$$\begin{aligned} \text{Then, } & f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ & \leq \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(y) \\ & = \lambda(f(x) + g(x)) + (1-\lambda)(f(y) + g(y)) \quad [\because f \text{ convex, } g \text{ linear}] \end{aligned}$$

Hence, $f+g$ is convex. ... Q

- Lemma: - Let $f: A \rightarrow B$, $g: B \rightarrow \mathbb{R}$
such that A, g range of f are convex
sets and f is a linear map, and
 g is convex (resp. concave). Then
 $g \circ f$ is convex (resp. concave).

Proof: - Choose $\lambda \in [0,1]$ and let
 $x, y \in A$. We show the proof
for when g is convex. The proof
for g concave follows similarly.

$$g \circ f(\lambda x + (1-\lambda)y)$$

$$= g(\lambda f(x) + (1-\lambda)f(y)) \quad [\because f \text{ is linear}]$$

$$\leq \lambda g(f(x)) + (1-\lambda)g(f(y)) \quad [\because g \text{ is convex}].$$

Hence, gof is convex.

... \square

Theorem :- Let $X, Y \sim P_{XY} = P_X \cdot P_{Y|X}$.

Then, with $P_{Y|X}$ fixed, $I(X; Y)$ is a concave function of P_X . Also,

with P_X fixed, $I(X; Y)$ is a convex function of $P_{Y|X}$.

Proof:- First let us fix $P_{Y|X}$.

Then, note that

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

$$= H(Y) - \sum_{x \in X} P_X(x) H(Y|X=x)$$

Now, note that

$$P_Y(\cdot) = \sum_{x \in X} P_{XY}(x, \cdot) = \sum_{x \in X} P_X(x) \cdot P_{Y|X}(\cdot | x).$$

Thus, with $P_{Y|X}$ fixed, $H(Y)$ is a linear function of P_X .

Next, we have already shown that $H(Y)$ is a concave function of P_Y .

Thus, by the previous lemma,

$H(Y)$ is a concave function of P_X .

On the other hand, fixing $P_{Y|X}$,

$H(Y|X=x)$ is fixed.

Thus, $H(Y|X) = \sum_{x \in X} P_X(x) \cdot H(Y|X=x)$

is a linear function of P_X .

Hence, by one of the previous lemmas,

$I(X;Y) = H(Y) - H(Y|X)$ is
a concave function of P_x if
 $P_{Y|X}$ is fixed.

Alternatively, assume that P_x is
fixed.

Then, $I(X;Y) = D(P_{XY} \parallel P_x \cdot P_Y)$

So, let $P_{Y|X}^{(1)}, P_{Y|X}^{(2)}$ be any two
conditional pmf of Y given X .

$$\text{let } P_{XY}^{(1)} = P_x \cdot P_{Y|X}^{(1)}$$

$$P_{XY}^{(2)} = P_x \cdot P_{Y|X}^{(2)}$$

$$\text{Also, } P_Y^{(1)} = \sum_{x \in X} P_x(x) \cdot P_{Y|X}^{(1)}(\cdot|x)$$

$$P_Y^{(2)} = \sum_{x \in X} P_x(x) \cdot P_{Y|X}^{(2)}(\cdot|x)$$

Now, pick any $\lambda \in [0, 1]$.

Now, consider the conditional pmf

$P_{Y|X}^* = \lambda P_{Y|X}^{(1)} + (1-\lambda) P_{Y|X}^{(2)}$, and note that it is well defined, i.e. $P_{Y|X}^*$ is indeed a conditional pmf.

$$\begin{aligned} \text{Then, } P_{XY}^* &= P_X \cdot P_{Y|X}^* \\ &= \lambda P_X \cdot P_{Y|X}^{(1)} + (1-\lambda) \cdot P_X \cdot P_{Y|X}^{(2)} \\ &= \lambda P_{X,Y}^{(1)} + (1-\lambda) \cdot P_{X,Y}^{(2)} \end{aligned}$$

On the other hand

$$\begin{aligned} P_Y^*(\cdot) &= \sum_{x \in X} P_X(x) \cdot P_{Y|X}^*(\cdot|x) \\ &= \lambda P_Y^{(1)} + (1-\lambda) P_Y^{(2)} \end{aligned}$$

$$\text{Thus, } P_X \cdot P_Y^*(\cdot) = \lambda P_X P_Y^{(1)} + (1-\lambda) P_X \cdot P_Y^{(2)}.$$

Thus, by convexity of relative entropy,
we have

$$\begin{aligned} D(P_{xy}^* \parallel P_x P_y^*) \\ \leq \lambda D(P_{xy}^{(1)} \parallel P_x P_y^{(1)}) \\ + (1-\lambda) D(P_{xy}^{(2)} \parallel P_x P_y^{(2)}) \end{aligned}$$

Hence, mutual information is a
convex function of $P_{y|x}$ given P_x is
fixed

..... B

• Lemma (Data Processing Inequality): -

Let X, Y, Z form a Markov

Chain, which we denote by $X-Y-Z$.

Then, (i) $I(X;Z) \leq I(X,Y)$

(ii) $I(X;Z) \leq I(Y;Z)$.

We have an equality in (i) iff
 $X-Z-Y$, in (ii) iff $Y-X-Z$

Proof :- $I(X;YZ) = I(X;Y)$
 + $I(X;Z|Y)$
 = $I(X;Y)$ $\because X-Y-Z$

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

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

Note that we have equality

iff $I(X;Y|Z)=0$, i.e.
 $X-Z-Y$.

Similarly,

$$\begin{aligned} I(XY;Z) &= I(Y;Z) + I(X;Z|Y) \\ &= I(Y;Z) \quad [\because X - Y - Z] \end{aligned}$$

Also, $I(XY;Z)$

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

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

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

i.e., $Y - X - Z$.

. 

Channel coding for discrete memoryless

channels

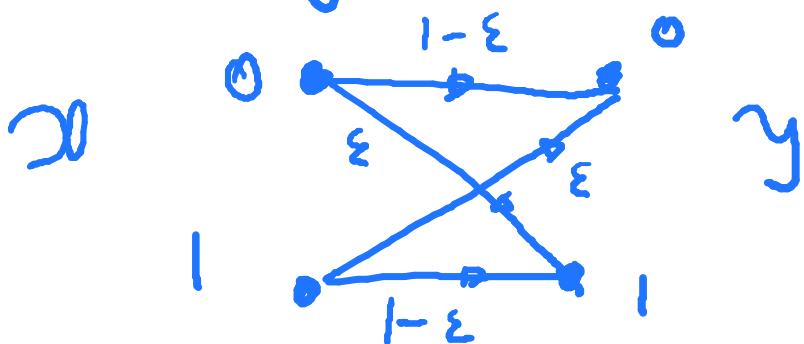
* Definition:- A discrete memoryless channel with ^{discrete} input alphabet X , ^{discrete} output alphabet Y , and a channel transition probability

$P_{Y|X}$ is a system, which on given an input $x^n \in \mathcal{X}^n$, for any $n \geq 1$, outputs a random y^n in \mathcal{Y}^n s.t.

$$Y^n \sim \prod_{i=1}^n P_{Y|X=x_i}$$

So, essentially, a channel is like a blackbox which given a string of length n , outputs a random string of length n , where i th letter is randomly sampled according to $P_{Y|X=x_i}$, independently of all other letters.

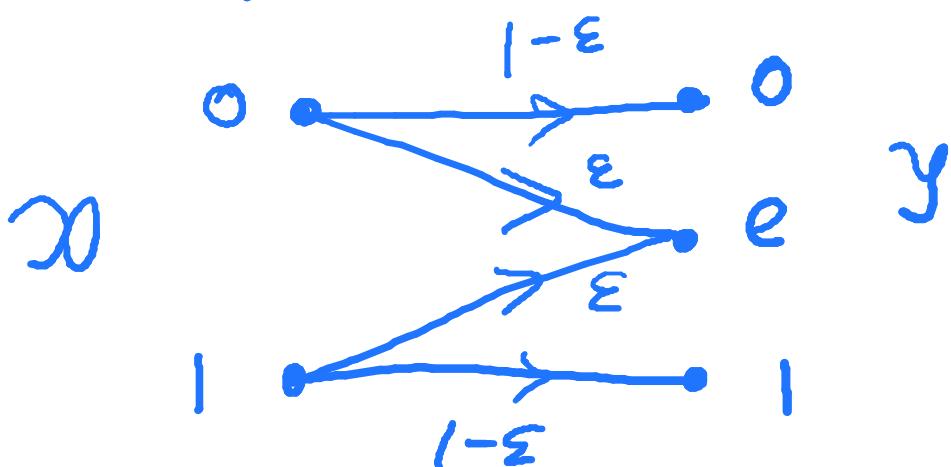
e.g. 1. Binary symmetric channel



$$\mathcal{X} = \mathcal{Y} = \{0, 1\}, P_{Y|X} = \begin{bmatrix} 0 & 1 \\ 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}.$$

ε is called the crossover probability of the binary symmetric channel. In shorthand, we denote this channel by $BSC(\varepsilon)$.

2. Binary erasure channel (BEC(ε))



$$\mathcal{X} = \{0, 1\}$$

$$\mathcal{Y} = \{0, 1, e\}$$

$$P_{Y|X} = \begin{matrix} & \begin{matrix} 0 & 1 & e \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{bmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{bmatrix} \end{matrix}$$

* Definition (Rate) :- A number $R \geq 0$ is said to be the achievable communication rate over a DMC with input alphabet \mathcal{X} , output alphabet \mathcal{Y} , and channel transition probability $P_{Y|X}$ if there exists a sequence of encoder-decoder pairs $f^{(n)}: \mathcal{M}_n \rightarrow \mathcal{X}^n$, $g^{(n)}: \mathcal{Y}^n \rightarrow \mathcal{M}_n$ such that,

- $\liminf_{n \rightarrow \infty} \frac{1}{n} \log |\mathcal{M}_n| = R$
- $\lim_{n \rightarrow \infty} P(g^{(n)}(f^{(n)}(w)) \neq w) = 0$,
where $w \sim \text{unif} \{\mathcal{M}_n\}$.

Remark (Important):-

Note that in the definition of rate, we only want the probability of error averaged over $w \sim \text{unif}\{\mathcal{M}_n\}$ to go down to 0. We can strengthen the definition by changing (ii) to demand that $\lim_{n \rightarrow \infty} \max_{w \in \mathcal{M}_n} P(g^{(n)}(f^{(n)}(w)) \neq w) = 0$.

We will show later that this modification does not change the expression for capacity.

* Definition (Channel capacity) :-

The capacity of a discrete memoryless channel is the supremum of all achievable communication rates.

• Theorem (Channel coding theorem) :-

For any DMC with input alphabet X , output alphabet Y , and channel transition probability $P_{Y|X}$, the channel capacity is given by,

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

Remarks :-

- (i) Notice that fixing P_x , the joint distribution P_{XY} is fixed as $P_{Y|X}$ is already fixed. So by fixing P_x , we can compute $I(X;Y)$.
- (ii) The channel coding theorem links an operational quantity C to the expression $\max_{P_X} I(X;Y)$.

Capacity is not defined as $\max_{P_X} I(X;Y)$, it is proved to be equal to it.

(iii) Note that we have used \max_{P_X} and
not \sup_{P_X} . This is justified as follows.

Let P_X be the set of all pmfs
over X .

Step 0:- Show that (P_X, TV) and (P_X, l_2)
are 'equivalent' metric spaces, i.e.
open sets in (P_X, TV) are open sets in (P_X, l_2)
and vice versa. This immediately implies
the same for closed sets since complements
of open sets are closed and vice
versa. To do so, use the following
easy to prove results

$$(i) \quad l_2(P, Q) \leq l_1(P, Q) \leq \sqrt{|X|} l_2(P, Q)$$

$$(ii) \quad TV(P, Q) = \frac{1}{2} l_1(P, Q)$$

Note that the equivalence of open sets
imply equivalence of compact sets as well.

Step 1 :- Prove that P_X is compact under $TV(\cdot, \cdot)$.

- Let $P_X^{(n)}, n \geq 1$, be a

Cauchy sequence of pmfs from P_X , w.r.t TV metric.

Then, for every $x \in X$, show

that $P_X^{(n)}(x), n \geq 1$, is a

Cauchy sequence in \mathbb{R} . Then,

let $P_X^*(x) = \lim_{n \rightarrow \infty} P_X^{(n)}(x)$,
 $\forall x \in X$.

It is then easy to see that

$P_X^*(x) \geq 0, \forall x$

and $\sum_{x \in X} P_X^*(x) = 1$.

Thus, P_X^* is a pmf.

Next show that $P_X^{(n)}$ converges to

P_x^* w.r.t. TV.

Thus, as $P_x^* \in \mathcal{P}_x$, \mathcal{P}_x is closed under TV

- By step 0, \mathcal{P}_x is closed under l_2 .

- \mathcal{P}_x is of course bounded

- By Heine-Borel theorem

\mathcal{P}_x is compact under l_2 .

[This is why we needed Step 0.]

- Thus \mathcal{P}_{x_0} is compact under TV.

Step 2 :- Show that $I(x; Y)$

is a continuous function

of P_x with $P_{Y|X}$ fixed. To see this

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

$$\text{and } P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) \cdot P_{Y|X}(y|x)$$

$$\text{and } H(Y|X) = \sum_{x \in \mathcal{X}} P_X(x) \cdot H(Y|X=x).$$

Then by continuity of entropy,
the fact that linear maps between
finite dimensional spaces are
continuous, and that
composition of continuous maps
are continuous, we get that
 $I(X;Y)$ is a continuous
function of P_X .

Step 3:- Since $I(X;Y)$ is
continuous and P_X is
compact, the set

$\mathcal{R} \stackrel{\text{def}}{=} \{r : \exists P_x \in \mathcal{P}_x \text{ s.t. } I(X; Y) = r\}$ is
also compact. Hence it is
also closed. Thus,

$$\exists r^* \in \mathcal{R}$$

$$r^* = \sup_{r \in \mathcal{R}} r.$$

Hence, $\sup_{P_x} I(X; Y) = \max_{P_x} I(X; Y)$.

iii) Since $I(X; Y)$ is a concave
function of P_X , one can use
common optimisation techniques
to find capacity. For example,
see the Blahut - Arimoto
algorithm.

* Jointly typical sequences

Define the set of jointly typical sequences

as,

$$A_{\varepsilon, xy}^{(n)} \triangleq \left\{ (x^n, y^n) : \left| \frac{1}{n} \log \frac{1}{P_x(x^n)} - H(X) \right| < \varepsilon, \right.$$

$$\left| \frac{1}{n} \log \frac{1}{P_y(y^n)} - H(Y) \right| < \varepsilon, \right.$$

$$\left. \left| \frac{1}{n} \log \frac{1}{P_{XY}(x^n, y^n)} - H(X, Y) \right| < \varepsilon \right\}.$$

• Lemma (AEP for joint typicality) :-

(i) Let $(X^n, Y^n) \sim \text{i.i.d } P_{XY}$. Then

$$P_{XY}(A_{\varepsilon, XY}^{(n)}) \rightarrow 1.$$

$$(ii) |A_{\varepsilon, XY}^{(n)}| \leq 2^{n(H(X, Y) - \varepsilon)}$$

(iii) Let $(\tilde{X}^n, \tilde{Y}^n) \sim \text{i.i.d } P_X P_Y$. Then,

$$(1-\varepsilon) \cdot 2^{-n(I(X;Y)+3\varepsilon)} \leq_n P_n((\tilde{X}^n, \tilde{Y}^n) \in A_{\varepsilon, XY}^{(m)})$$

$$\leq 2^{-n(I(X;Y)-3\varepsilon)}.$$

Proof :- (ii) Note that

$$E_{P_X} \left[\log \frac{1}{P_X(x)} \right] = H(X)$$

$$E_{P_Y} \left[\log \frac{1}{P_Y(y)} \right] = H(Y)$$

$$E_{P_{XY}} \left[\log \frac{1}{P_{XY}(x,y)} \right] = H(X,Y)$$

Now, fix any $\varepsilon > 0$.

Then, by WLLN,

$$P_{XY} \left(\left| \frac{1}{n} \log \frac{1}{P_X(x^n)} - H(X) \right| \geq \varepsilon \right) \leq_n \frac{\varepsilon}{3}$$

$$P_{XY} \left(\left| \frac{1}{n} \log \frac{1}{P_Y(y^n)} - H(Y) \right| \geq \varepsilon \right) \leq_n \frac{\varepsilon}{3}$$

$$P_{XY} \left(\left| \frac{1}{n} \log \frac{1}{P_{XY}(x^n, y^n)} - H(X, Y) \right| \geq \varepsilon \right) \leq_n \frac{\varepsilon}{3}$$

Thus, by union bound,

$$P_{XY}(A_{\varepsilon, XY}^{(n)}) \leq \frac{\varepsilon}{3} \cdot 3 = \varepsilon$$

Since choice of $\varepsilon > 0$ was arbitrary,
we have,

$$P_{XY}(A_{\varepsilon, XY}^{(n)}) \rightarrow 1.$$

(ii) $| \geq P_{XY}(A_{\varepsilon, XY}^{(n)})$

$$= \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon, XY}^{(n)}}} P_{XY}(x^n, y^n)$$

$$\geq \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon, XY}^{(n)}}} 2^{-n(H(X,Y) + \varepsilon)}$$

$$= |A_{\varepsilon, XY}^{(n)}| \cdot 2^{-n(H(X,Y) + \varepsilon)}$$

$$\Rightarrow |A_{\varepsilon,xy}^{(n)}| \leq 2^{n(H(x,y) + \varepsilon)}$$

$$(iii) P_n(\tilde{X}^n, \tilde{Y}^n \in A_{\varepsilon,xy}^{(n)})$$

$$= \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon,xy}^{(n)}}} P_n(\tilde{X}^n = x^n, \tilde{Y}^n = y^n)$$

$$= \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon,xy}^{(n)}}} P_x(x^n) \cdot P_y(y^n)$$

$$\leq \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon,xy}^{(n)}}} 2^{-n(H(x) - \varepsilon)} \cdot 2^{-n(H(y) - \varepsilon)}$$

$$\leq |A_{\varepsilon,xy}^{(n)}| \cdot 2^{-n(H(x) + H(y) - 2\varepsilon)}$$

$$\leq |A_{\varepsilon,xy}^{(n)}| \cdot 2^{-n(H(x) + H(y) - H(x,y) - 3\varepsilon)}$$

[By part (ii)]

$$= 2^{-n(I(x;y) - 3\varepsilon)}$$

$$P_n(\tilde{X}^n, \tilde{Y}^n \in A_{\varepsilon, XY}^{(n)})$$

$$= \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon, XY}^{(n)}}} P_X(x^n) \cdot P_Y(y^n)$$

$$\geq \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon, XY}^{(n)}}} 2^{-n(H(X) + H(Y) + 2\varepsilon)}$$

$$= |A_{\varepsilon, XY}^{(n)}| \cdot 2^{-n(H(X) + H(Y) + 2\varepsilon)} \quad \dots \quad \textcircled{1}$$

Now, by part (i)

$$P_{XY}(A_{\varepsilon, XY}^{(n)}) \geq_n 1 - \varepsilon$$

$$\Rightarrow \sum_{\substack{(x^n, y^n) \in \\ A_{\varepsilon, XY}^{(n)}}} P_{XY}(x^n, y^n) \geq_n 1 - \varepsilon$$

$$\Rightarrow |A_{\varepsilon, XY}^{(n)}| \cdot 2^{-n(H(X, Y) - \varepsilon)} \geq_n 1 - \varepsilon$$

Plugging this in ① we get

$$\Pr(\tilde{x}^n, \tilde{y}^n \in A_{\varepsilon, XY}^{(n)}) \geq_n (1-\varepsilon) \cdot 2^{-n(H(X)+H(Y)+2\varepsilon)}$$
$$= (1-\varepsilon) \cdot 2^{-n(I(X;Y)+3\varepsilon)}$$

* What is a probabilistic method?

Idea:- Suppose you want to prove a property of certain objects. Suitable choose a probability distribution over those objects. Show that the property doesn't hold with probability strictly less than one.

This implies that there exist an instance of that object for which the property must hold.

We shall use this idea to prove the achievability part of the channel coding theorem.

We first illustrate this technique with a simple problem.

• Definition:- A t -uniform hypergraph $H = (V, E)$ consists of a finite set V , and a set of its t -sized subsets E .

- Definition :- Given a hypergraph $H = (V, E)$, we say that it has a two-colouring if there exists a map $f: V \rightarrow \{0, 1\}$ s.t. $\forall e \in E$, $\exists i, j \in e$ s.t. $f(i) \neq f(j)$.

- Lemma :- There exists a two-colouring of a t -uniform hypergraph $H = (V, E)$ if $|E| \leq 2^{t-1}$.

Proof :- Colour every vertex in V with 0 and 1 uniformly, and independently of all other vertices.

Fix $e \in E$. Then the probability that every xee has the same colour is 2^{-t+1} , since $|R|=t$.

Then, by union bound, the random assignment is not a two-colouring with probability at most

$$|E| \cdot 2^{-t+1} < 1 \quad (\because |E| \leq 2^{-t+1})$$

Thus, there exists a two-colouring of H .



* Achievability proof : - Any $R < \max_{P_x} I(X;Y)$ is achievable

Fix any P_x , any $\delta > 0$, and any $R < I(X;Y) - 3\delta$

Fix n , and fix $M_n = \{1, 2, \dots, \lceil 2^{nR} \rceil\}$, and
fix any $0 < \varepsilon < \delta$

We define a random codebook ($\text{or } f^{(n)}$)
as follows.

For any $w \in M_n$, assign $f^{(n)}(w) = x^{(n)}(w)$
with probability $\prod_{i=1}^n P_X(x_i(w))$.

We shall interchangeably refer to

the encoding map as a codebook.

We shall use ℓ to denote a random
codebook, and for any instance of a
codebook $c = (x_c^n(1), \dots, x_c^n(\lceil 2^{nR} \rceil))$,

note that $P_n(\ell = c) = \prod_{w \in M_n} \prod_{i=1}^n P_X(x_{c,i}(w))$.

Given a codebook (\mathcal{C} or $f^{(n)}$), next we specify the decoding rule as follows.

Let the decoder sees $y^n \in \mathcal{Y}^n$.

- The decoder decodes to $w \in M_n$, if
$$(f^{(n)}(w), y^n) \in A_{\mathcal{C}, xy}^{(n)} \text{ and } \exists w' \neq w$$
s.t.
$$(f^{(n)}(w'), y^n) \in A_{\mathcal{C}, xy}^{(n)}$$
- If the above condition is not met, then the decoder declares decoding failure.

[Note that the decoder will of course know the encoding map $f^{(n)}$ / the codebook \mathcal{C} , and hence it can perform the decoding mentioned above.]

Let $\Sigma \triangleq \{ g^{(n)}(f^{(n)}(w)) \neq w \}$ be the error event.

Note here that Σ has three sources of "randomness"

$$\text{i) } w \sim \text{unif}(M_n)$$

ii) The channel

iii) The choice of codebook \mathcal{C} , i.e., the choice of $(f^{(n)}, g^{(n)})$.

Now, note that by law of total probability

$$\begin{aligned} \Pr_n(\Sigma) &= \sum_c \Pr_n(c) \cdot \Pr_n(\Sigma | c) \\ &= \mathbb{E}_c [\Pr_n(\Sigma | c)] \end{aligned}$$

We shall show that

$$\mathbb{E}_e [P_n(\varepsilon | e)] \leq_n \varepsilon$$

This would then imply that the sequence $\min_c P_n(\varepsilon | e=c) \leq_n \varepsilon$. (Note that this is indeed a minimum as the set of all possible codebooks is finite)

$$\text{det } C_n^* \in \arg \min_c P_n(\varepsilon | e=c).$$

Noting that $\varepsilon > 0$ is arbitrary and C_n^* is independent of ε , we have a sequence of codebooks C_n^* s.t.

$$P_n(\varepsilon | e=C_n^*) \leq_n \varepsilon \Rightarrow \lim_{n \rightarrow \infty} P_n(\varepsilon | e=C_n^*) = 0$$

The rate of that codebook C_n^* will be

$$R \leq \frac{\log |M_n|}{n} \leq_n R + \frac{1}{n} \quad \left[\because \lceil 2^{nR} \rceil \leq 2^{nR} + 1 \text{ and } \log(2^{nR} + 1) \leq nR + 1 \right]$$

$$\text{And hence, } \lim_{n \rightarrow \infty} \frac{\log |M_n|}{n} = R.$$

Now, choosing the appropriate

P_x that maximises $I(X; Y)$, and $\delta > 0$, we

can choose any $R \leq \max_{P_x} I(X; Y)$, which will complete the achievability proof.

Thus, it remains to show that

$$E_e[P_n(\epsilon | e)] \leq_n \epsilon$$

Define $P_{e,w} \doteq P_n(\epsilon | W=w)$

Thus, $E_e[P_n(\epsilon | e)]$

$$= \sum_c P_n(e=c) \cdot P_n(\epsilon | e=c)$$

$$= \sum_c \sum_{w \in M_n} P_n(e=c) \cdot \frac{1}{|M_n|} \cdot P_n(\epsilon | e=c, W=w)$$

[By law of total probability]

and noting that the codebook/fin

is picked independent of the realization
of w]

$$= \sum_{w \in M_n} \frac{1}{|M_n|} \sum_c P_n(C=c) \cdot P_n(E|C=c, W=w)$$

..... ①

Next, we claim that for any $w \neq 1$,

$$\begin{aligned} & \sum_c P_n(C=c) \cdot P_n(E|C=c, W=w) \\ &= \sum_c P_n(C=c) \cdot P_n(E|C=c, W=1) \end{aligned}$$

To see this, fix $w \neq 1$, and observe

that every codebook C , one can

obtain the codebook C' s.t.

$$x_C^n(w') = x_{C'}^n(w'), \quad w' \neq 1, w$$

$$x_C^n(1) = x_{C'}^n(w)$$

$$x_C^n(w) = x_{C'}^n(1)$$

It is easy to see

- $P_n(\ell=c) = P_n(\ell=c')$ by the way the probability of codebook were chosen
- $P_n(\varepsilon | w=1, \ell=c') = P_n(\varepsilon | w=w, \ell=c)$ by symmetry.

Thus, $\sum_c P_n(\ell=c) \cdot P_n(\varepsilon | w=w, \ell=c)$

$$= \sum_c P_n(\ell=c) \cdot P_n(\varepsilon | w=1, \ell=c).$$

Plugging this in ①, we get

$$\begin{aligned} \mathbb{E}_e [P_n(\varepsilon | e)] &= \sum_w \frac{1}{|M_n|} \cdot \sum_c P_n(\ell=c) \cdot \\ &\quad P_n(\varepsilon | w=1, \ell=c) \\ &= \sum_c P_n(\ell=c) \cdot P_n(\varepsilon | w=1, \ell=c) \end{aligned}$$

$$= \Pr_n(\mathcal{E} | W=1) \dots \textcircled{2}$$

Thus, we simply need to show that

$$\Pr_n(\mathcal{E} | W=1) \leq_n \varepsilon$$

[Note that the 'randomness' left in \mathcal{E} conditioned on $W=1$ are due to

- Random codebook \mathcal{C} / Choice of $f^{(n)}$
- The channel]

Now, given $W=1$, $\mathcal{E} = \mathcal{E}_1^c \cup \left(\bigcup_{w \neq 1} \mathcal{E}_w \right)$,

$$\text{where } \mathcal{E}_w = \left\{ (x_e^n(w), Y^n) \in A_{\delta, XY}^{(n)} \right\}$$

Now, by union bound,

$$\Pr_n(\mathcal{E} | W=1) \leq \Pr_n(\mathcal{E}_1^c | W=1)$$

$$+ \sum_{w \neq 1} \Pr_n(\mathcal{E}_w | W=1)$$

Given $W=1$, observe that

$$(x_e^n(1), Y^n) \sim P_{XY}^n.$$

Thus, by the joint AEP,

$$P_n(\mathcal{E}_1 | W=1) \leq \frac{\epsilon}{2}$$

Again, observe that when conditioned

on $W=1$, for any $w \neq 1$, Y^n depends
on $x_e^n(1)$ but is independent of
 $x_e^n(w)$. Also, by construction of ℓ ,

$$(x_e^n(w), Y^n) \sim P_X \cdot P_Y$$

Thus, by joint AEP,

$$-n(I(X; Y) + 3\epsilon)$$

$$P_n(\mathcal{E}_w | W=1) \leq 2$$

Thus, $E_\epsilon [P_n(\epsilon | \epsilon)]$

$$\leq \frac{\epsilon}{2} + |M_n| \cdot 2^{-n(I(x;Y)-3\epsilon)}$$

$$\leq \frac{\epsilon}{2} + 2^{-n(I(x;Y)-3\epsilon)}$$

$$+ 2^{-n(I(x;Y)-R-3\epsilon)}$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} \quad \left[\because R < I(x;Y) - 3\delta \right.$$

$$< I(x;Y) - 3\epsilon$$

$$= \epsilon. \quad \text{by choice}$$

..... \blacksquare

Corollary:- By changing the definition

of capacity to include $\lim_{n \rightarrow \infty} \min_{w \in M_n} P_n(g^{(n)}(f^{(n)}(w)) \neq w) = 0$

instead of $\lim_{n \rightarrow \infty} P_n(g^{(n)}(f^{(n)}(w)) \neq w) = 0$, $w \sim \text{unif}(M_n)$,

we still have any $R < \max_x I(x;Y)$ to be achievable.

Proof:- In the proof of the achievability part, we showed that for any P_x , and any $\delta > 0$, and any $R < I(x; Y) - 3\delta$, there exists a sequence of codebooks, C_n , which for any $0 < \epsilon < \delta$ satisfies ..

$$P_n(\epsilon | e = c_n) \leq \epsilon$$

$$\text{and } \lim_{n \rightarrow \infty} \frac{\log(M_n)}{n} = R$$

We shall now construct a new message set $M'_n = \{1', 2', \dots, [2^{nR}]'\}$

and new code-book c'_n not depending on

$$\epsilon, \text{ s.t. } \max_{w \in M'_n} P_n(\epsilon | e = c'_n, w = w) \leq 2\epsilon.$$

Noting that $\epsilon > 0$ was arbitrary,

this would imply $\lim_{n \rightarrow \infty} \max_{w \in M'_n} P_n(f^{(n)}(f^{(n)}(w)) = w) = 0$,

And as

$$R - \frac{1}{n} \leq \frac{\log |M_n'|}{n} \leq R - \frac{1}{n} + \frac{1}{n},$$

i.e., $\lim_{n \rightarrow \infty} \frac{\log |M_n'|}{n} = R$

Thus, R will be an achievable rate according to the max probability of error definition,

By appropriately choose

P_x that maximises $I(X;Y)$, and $\delta > 0$,

any $R < \max_{P_x} I(X;Y)$ can thus

be shown to be achievable and

hence $\max_{P_x} I(X;Y) \leq C$ according

to the max probability def".

It remains to construct c'_n and M'_n .

Relabel the elements of M_n s.t.

$$\Pr_n(E | \ell = c_n, W=1) \leq \Pr_n(E | \ell = c_n, W=2) \leq \dots$$
$$\leq \Pr_n(E | \ell = c_n, W=\lceil 2^{n\kappa} \rceil)$$

and define $M'_n = \{1', 2', \dots, \lceil 2^{n\kappa} \rceil'\}$

where $w' = w$.

Define $c'_n = x_{c_n}^{-n}(1'), \dots, x_{c_n}^{-n}(\lceil 2^{n\kappa} \rceil')$.

where $x_{c_n}^{-n}(w') = x_{c_n}^{-n}(w)$.

We claim that

$$\Pr_n(E | \ell = c'_n, W=w') \leq_n 2\varepsilon, \forall w' \in M'_n,$$

and thus we will have

$\max_{w' \in M'_n} \Pr_n(E | \ell = c'_n, W=w') \leq_n 2\varepsilon$, and the proof follows.

To see this, observe that

$$q_E \geq_n \sum_{w \in M_n} \frac{2}{|M_n|} P_n(E | e=c_n, W=w)$$

$$= \sum_{w \in M_n} \frac{1}{|M_n'|} P_n(E | e=c_n', W=w)$$

$$\geq \sum_{w > \frac{[2^{nR}]}{2}} \frac{1}{|M_n'|} P_n(E | e=c_n', W=w)$$

$$\geq \sum_{w > \frac{[2^{nR}]}{2}} \frac{1}{|M_n'|} \max_{w' \in M_n'} P_n(E | e=c_n', W=w')$$

[By construction]

$$= \max_{w' \in M_n'} P_n(E | e=c_n', W=w'),$$

as required. $\therefore \emptyset$

* Proof of converse :-

Any $R > \max_{P_x} I(X; Y)$ is not achievable

• Lemma (Fano's inequality) :-

Let X, \hat{X} be random variables on \mathcal{X} and let $P_r(X \neq \hat{X}) = p$. Then,

$$H(X|\hat{X}) \leq \log |\mathcal{X}| \cdot p + h(p)$$

Proof:- Let $Y = 1\{X \neq \hat{X}\}$

$$H(X|\hat{X}) \leq H(X, Y|\hat{X})$$

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

$$\leq H(Y) + P(Y=1) \cdot H(X|\hat{X}, Y=1) \\ + P(Y=0) \cdot H(X|\hat{X}, Y=0)$$

$$\leq h(p) + p \cdot \log |\mathcal{X}| + 0$$

$$[\because H(X|\hat{X}, Y=1) \leq H(X) \leq \log |X|]$$

$$H(X|\hat{X}, Y=0) = 0 \quad \because X = \hat{X} \text{ if } Y=0]$$

. . . . B

• Lemma :- Let X^n, Y^n be such that

$(X_{\neq i}, Y_{\neq i}) - X_i - Y_i$. Then,

$$I(X^n; Y^n) \leq \sum_{i=1}^n I(X_i; Y_i), \text{ with}$$

equality iff Y_1, \dots, Y_n are independent.

Proof :-

$$I(X^n; Y^n) = H(Y^n) - H(Y^n/X^n)$$

$$= \sum_{i=1}^n \left[H(Y_i/Y^{i-1}) - H(Y_i/X_i, X_{\neq i}, Y^{i-1}) \right]$$

$$\leq \sum_{i=1}^n \left[H(Y_i) - H(Y_i/X_i) \right] \quad [\text{By chain rule}]$$

$$[\because Y_i - X_i - (X_{\neq i}, Y^{i-1})]$$

$$= \sum_{i=1}^n I(X_i; Y_i)$$

Notice that we have equality iff

$$H(Y_i | Y^{i-1}) = H(Y_i) \quad \forall i \in [n], \text{ i.e.}$$

iff Y_1, \dots, Y_n are independent.

..



Remark:- In a channel, if X^n is sent,
 $X_{\neq i}, Y_{\neq i} - X_i - Y_i$ condition is met.

We now begin the proof of the converse part:

Firstly note that if R is an achievable rate according to the maximum probability error definition, it is by default an achievable rate according to an achievable probability of error definition.

Thus, we will only prove any

$R > \max_{P_X} I(X; Y)$ is not achievable according to the average probability of error definition.

Fix a sequence of encoder-decoder pairs $f^{(n)} - g^{(n)}$ of rate R , i.e., $\liminf_{n \rightarrow \infty} \frac{1}{n} \log |M_n| = R$

s.t. $\lim_{n \rightarrow \infty} \Pr(W + g^{(n)}(f^{(n)}(W))) = 0$.

Fix any $\frac{1}{2} > \varepsilon > 0$, and let $\hat{W} = g^{(n)}(f^{(n)}(W))$.

Then, we must have

$$\Pr_n(W \neq \hat{W}) \leq \varepsilon.$$

Then, by Fano's inequality

$$H(W|\hat{W}) \leq h(\varepsilon) + \varepsilon \cdot \log |M_n|.$$

($\because h(\cdot)$ is increasing in $[0, \frac{1}{2}]$)

$$\text{Now, } H(W|\hat{W})$$

$$= H(W) - I(W; \hat{W})$$

$$\geq \log |M_n| - I(X^n; Y^n)$$

$(\because W = X^n = Y^n = \hat{W},$

and hence by DPI

$$I(W; \hat{W}) \leq I(X^n; Y^n)$$

$$\geq \log |M_n| - \sum_{i=1}^n I(X_i; Y_i)$$

[By the lemma]

$$\geq \log |M_n| - n \cdot C' \quad [\text{where}]$$

$$C' \triangleq \max_{P_X} I(X; Y)$$

Thus, putting thing together, we have

$$\frac{1}{n} \log |M_n| - C' \leq \varepsilon \cdot \frac{\log |M_n|}{n} + \frac{1}{n} h(\varepsilon)$$

$$\Rightarrow \frac{1}{n} \log |M_n| \leq (C' + \frac{1}{n} h(\varepsilon)) \cdot \frac{1}{(1-\varepsilon)}$$

$$\Rightarrow R \leq \frac{C'}{1-\varepsilon} \quad [\because a_n \leq b_n \text{ and } b_n \rightarrow b \Rightarrow \liminf_{n \rightarrow \infty} a_n \leq b]$$

To see this,

$$\liminf_{n \rightarrow \infty} a_n = \sup_{k \geq 1} \inf_{n \geq k} a_n$$

$$\because a_n \leq b_n, \exists k^* \text{ s.t.}$$

$$a_n \leq b_n, \forall n \geq k^*$$

Also, $\inf_{n \geq k^*} a_n$ is an increasing sequence in \mathbb{R} .

Sequence is R .

$$\text{Thus } \sup_{k \geq 1} \inf_{n \geq k} a_n = \sup_{k \geq k^*} \inf_{n \geq k} a_n$$

$$\leq \sup_{k \geq k^*} \inf_{n \geq k} b_n \quad [\because a_n \leq b_n \quad \forall n \geq k^*]$$

$$= \liminf_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} b_n = b.$$

So, we have that for $\varepsilon > 0$, an achievable

$$R \text{ satisfies } R \leq \frac{C'}{1-\varepsilon}.$$

This thus implies $R < C' \stackrel{\text{def}}{=} \max_{P_X} I(X; Y)$
as required.

[To see why this is so, assume

$$R \geq C', \text{ i.e. } R = C' + \delta, \delta \geq 0.$$

Then, $R = C'\left(1 + \frac{\delta}{C'}\right).$

Choose ε s.t. $1 + \frac{\delta}{C'} = \frac{1}{1-\varepsilon}$, i.e.

$$\varepsilon = 1 - \frac{1}{1 + \frac{\delta}{C'}}. \text{ It is then easy to}$$

see that $\frac{1}{1 - \frac{\varepsilon}{2}} > 1 + \frac{\delta}{C'}$, and hence

$$R > \frac{C'}{1 - \frac{\varepsilon}{2}}, \text{ which violates what we derived.}]$$

* Channel capacity examples

- Lemma :- The capacity of BSC(ϵ) is $1 - h(\epsilon)$, and the distribution P_X achieving capacity is $B_{\epsilon}(\frac{1}{2})$.

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

$$\leq 1 - H(Y|X)$$

$$= 1 - P_n(X=0) \cdot H(Y|X=0)$$

$$- P_n(X=1) \cdot H(Y|X=1)$$

$$= 1 - h(\epsilon) \cdot [P_n(X=0) + P_n(X=1)]$$

$$= 1 - h(\epsilon).$$

Thus, it is enough to show that

$P_X = B_{\epsilon}(\frac{1}{2})$ satisfies $H(Y) = 1$.

This is obvious since

$$P_Y(1) = (1-\epsilon) \cdot P_X(1) + \epsilon \cdot P_X(0)$$

$$= \frac{1}{2} \cdot (1-\epsilon) + \frac{1}{2} \epsilon = \frac{1}{2} \quad \dots \quad \square$$

• Lemma :- The capacity of BEC(ε) is $1-\varepsilon$, and is achieved by $P_x = B_e(\frac{1}{2})$.

Proof:- Let $p = P_x(1)$ and thus
 $1-p = P_x(0)$.

$$\text{Then, } P_{X|Y}(0|0) = 1$$

$$P_{X|Y}(1|1) = 1$$

$$P_{X|Y}(0|e) = \frac{(1-p) \cdot \varepsilon}{\varepsilon \cdot (1-p+p)} = 1-p$$

$$P_{X|Y}(1|e) = p.$$

$$\text{Thus, } H(X|Y) = P_Y(e) \cdot h(p)$$

$$= \varepsilon \cdot (p + (1-p)) \cdot h(p)$$

$$= \varepsilon \cdot h(p)$$

$$\text{Thus, } I(X;Y) = H(X) - H(X|Y)$$

$$= (1-\varepsilon) h(p) \leq 1-\varepsilon.$$

Thus, $C \leq 1 - \varepsilon$, and with $P_x = B_2\left(\frac{1}{2}\right)$,
 $h(p) = 1$, and so, $I(X;Y) = 1 - \varepsilon$.

Thus, $C = 1 - \varepsilon$ \square

* Capacity with feedback

- Definition :- A real $R > 0$ is said to be an achievable rate for a DMC with feedback, if there exists a sequence of sets M_n , and a sequence of encoder-decoder maps $f_1^{(n)}, f_2^{(n)}, \dots, f_n^{(n)}, g^{(n)}$ where $f_i^{(n)}: M_n \times Y^{i-1} \rightarrow X, 1 \leq i \leq n$,

and $g^{(n)}: Y^n \rightarrow M_n$ satisfying

$$i) \liminf_{n \rightarrow \infty} \frac{1}{n} \log |M_n| = R$$

$$ii) \lim_{n \rightarrow \infty} P_n(\hat{W} \neq W) = 0,$$

where $W \sim \text{unif} \{M_n\}$ and

$$\hat{W} = g^{(n)}(f_1^{(n)}(W), f_2^{(n)}(W, Y_1), \dots, f_n^{(n)}(W, Y^{n-1}))$$

Remark :- Just as in the case without feedback, we can define achievable rate with feedback using maximum probability of error as opposed to expected probability of error. The main result, as in the previous case, does not change.

- Definition :- The capacity of a DMC with feedback, C_{fb} , is defined to be the supremum of all achievable rates with feedback.

Theorem:- For any DMC,

$$C = C_{fb}.$$

Proof:- Observe that any encoder for channel without feedback, is an encoder for channel with feedback, and hence any rate achievable without feedback is also achievable with feedback.

$$\text{Thus, } C_{fb} \geq C.$$

Now, let R be any achievable rate with feedback. Let $\rho = \Pr(\hat{W} \neq W)$.

Fix any $\epsilon > 0$, and hence,

$$\rho \leq \epsilon.$$

Thus, by Fano's inequality

$$H(W|\hat{W}) \leq f \cdot \log |M_n| + h(f) \\ \leq_n \varepsilon \cdot \log |M_n| + h(\varepsilon)$$

Also,

$$H(W|\hat{W}) = H(W) - I(W; \hat{W})$$

$$\geq \log |M_n| - I(W; Y^n)$$

$\left[\because W - Y^n - \hat{W} \text{ and } \text{DPI} \right]$

$$= \log |M_n| - H(Y^n)$$

$$+ \sum_{i=1}^n H(Y_i | Y^{i-1}, W)$$

$$= \log |M_n| - H(Y^n)$$

$$+ \sum_{i=1}^n H(Y_i | Y^{i-1}, W, X_i)$$

$\left[\because X_i = f_i^{(n)}(W, Y^{i-1}) \right]$

$$= \log |M_n| - H(Y^n) + \sum_{i=1}^n H(Y_i | X_i)$$

$\left[\because \text{By definition of channel } (W, Y^{i-1}) - X_i - Y_i \right]$

$$\geq \log |M_n| - \sum_{i=1}^n H(Y_i) + \sum_{i=1}^n H(Y_i | X_i)$$

$$= \log |M_n| - \sum_{i=1}^n I(X_i; Y_i)$$

$$\geq \log |M_n| - nC \quad [\because I(X_i; Y_i) \leq \max_{P_X} I(X; Y)]$$

Thus, combining we get

$$\frac{\log |M_n|}{n} \leq \left(C + \frac{h(\varepsilon)}{n} \right) \cdot \frac{1}{1-\varepsilon}$$

and thus, noting $R = \liminf_{n \rightarrow \infty} \frac{\log |M_n|}{n}$,

$$R \leq \frac{C}{1-\varepsilon}.$$

Since the choice of ε was arbitrary,

an achievable R satisfies $R < C$.

Thus, supremum of all achievable R

with feedback, $C_{fb} \leq C$, and this completes
the proof. ... B

• Remark:- Note that again the proof would go through if G_0 was defined using maximum probability of error as opposed to average probability of error.

* Joint source-channel coding theorem

Consider a DMS with alphabet

Σ and pmf P_v . Consider a DMC with input alphabet X , output alphabet Y , and channel transition probability $P_{Y|X}$.

• Definition:- A DMS with alphabet Σ
 and pmf P_V is said to be transmissible
 over a DMC with input alphabet X ,
 output alphabet Y , and channel
 transition probability $P_{Y|X}$, if for every
 $n \geq 1$, there exists encoder-decoder pairs
 $f^{(n)}: X^n \rightarrow V^n$, $g^{(n)}: Y^n \rightarrow V^n$, such that
 $\lim_{n \rightarrow \infty} P_n(V^n \neq g^{(n)}(f^{(n)}(V^n))) = 0$, where
 $V^n \sim \text{i.i.d. } P_V$.

• Theorem (Joint source-channel coding):-
 A DMS with pmf P_V is transmissible
 over a DMC with capacity C if
 $H \stackrel{def}{=} H(P_V) < C$, and not transmissible if $H > C$.

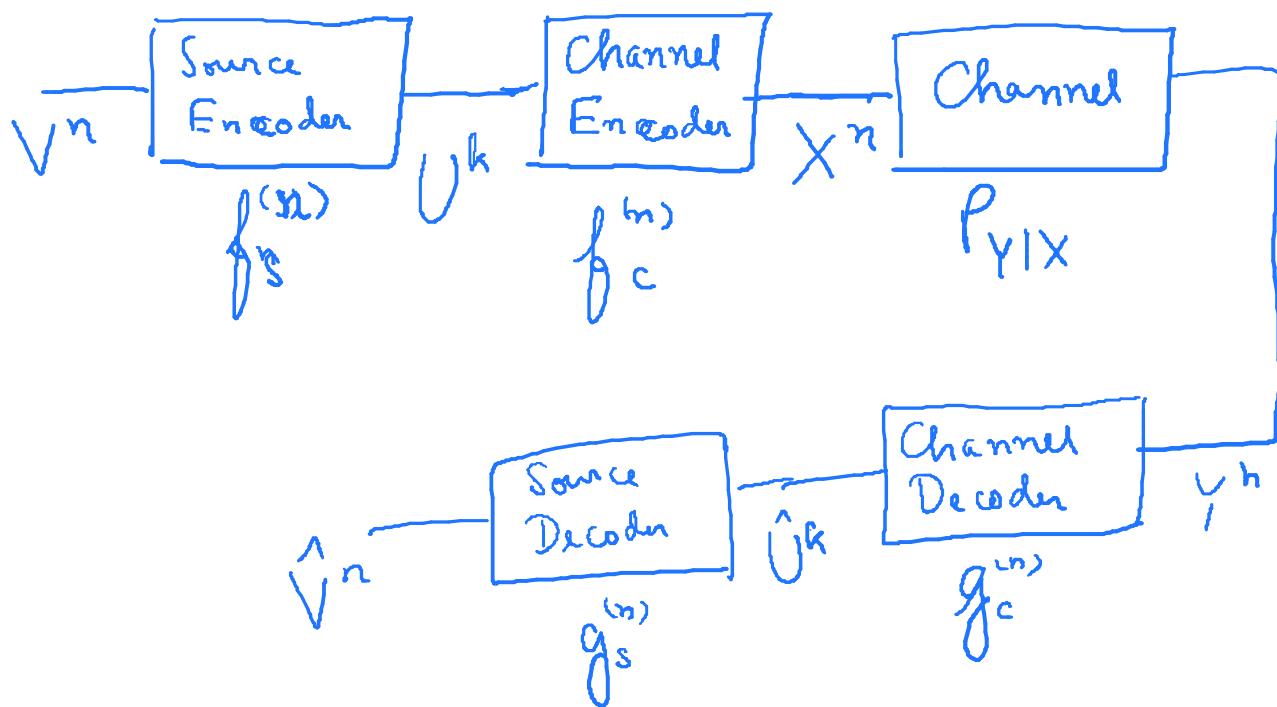
Proof:- Achievability :-

Fix an $\epsilon < \frac{C-H}{2}$. We shall

show the existence of an encoder-decoder pair $f^{(n)} - g^{(n)}$ s.t. $\Pr(V^n \neq g^{(n)}(f^{(n)}(V^n))) \leq \frac{\epsilon}{n}$.

This will prove the result.

The encoder-decoder pair is constructed by separating them into 'source coding' and 'channel coding' parts.



The source encoder $f_s^{(n)}: \mathcal{V}^n \rightarrow \{0,1\}^k$ is the familiar typical set encoder that encodes sequences in $A_{\epsilon,V}^{(n)}$ in one-one fashion to $\{0,1\}^k$, and the remaining sequences in $\mathcal{V}^n \setminus A_{\epsilon,V}^{(n)}$ to 0^k. By AEP,

we have, $|A_{\epsilon,V}^{(n)}| \leq 2^{n(H+\epsilon)}$, and hence we choose $k = \lceil n(H+\epsilon) \rceil$ to make $f_s^{(n)}$ well defined. Next, note that

$$\lim_{n \rightarrow \infty} \frac{k}{n} = H + \epsilon < C - \epsilon. \text{ Thus,}$$

$H + \epsilon$ is an achievable rate for the channel. So, choose $M_n = \{0, 1, \dots, 2^k\}$, and let $c_n = (x^n(1), x^n(2), \dots, x^n(2^k))$ be the

code book where $f_c^{(n)}: \{0,1\}^k \rightarrow \mathcal{X}^n$ defined as

$f_c^{(n)}(i) = \mathcal{X}^n(i)$, and $g_c^{(n)}: \mathcal{Y}^n \rightarrow \{0,1\}^k$ defined as the jointly typical decoder (as in

the achievability proof of the channel coding theorem), such that

$$\max_{i \in M_n} \Pr(g_c^{(n)}(f_c^{(n)}(i)) \neq i) \leq \frac{\epsilon}{2}. \quad \dots \textcircled{1}$$

Now, choose $f_s^{(n)} = f_c^{(n)} \circ f_s^{(n)}$ and

$$g_s^{(n)} = g_s^{(n)} \circ g_c^{(n)}. \text{ Then,}$$

$$\Pr(V^n \neq g_s^{(n)}(f_s^{(n)}(V^n)))$$

$$\leq \Pr(V^n \notin A_{\epsilon, V}^{(n)})$$

$$+ \Pr(V^n \neq g_s^{(n)}(f_s^{(n)}(V^n)) \mid V^n \in A_{\epsilon, V}^{(n)})$$

[Union bound +
law of total probability]

$$\leq_n \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad [\text{By ① and AEP}]$$

$$\leq \varepsilon.$$

Converse:- Now, let $f^{(n)}: \mathcal{V}^n \rightarrow \mathcal{W}^n$.

$g^{(n)}: \mathcal{Y}^n \rightarrow \mathcal{V}^n$ be a sequence of encoder-decoder pairs satisfying

$\Pr_n(V^n \neq g^{(n)}(f^{(n)}(V^n))) \rightarrow 0$. Fix any $\varepsilon > 0$.

Then, $\Pr_n(V^n \neq \hat{V}^n) \leq_n \varepsilon$, where

$$\hat{V} \triangleq g^{(n)}(f^{(n)}(V^n)).$$

$$\begin{aligned} \text{Now, } H &= \frac{H(V^n)}{n} \\ &= \frac{1}{n} H(V^n | \hat{V}^n) + \frac{1}{n} I(X^n; \hat{V}^n) \\ &\leq \frac{1}{n} n \varepsilon \log |\mathcal{V}| + h(\varepsilon) \\ &\quad + \frac{1}{n} I(X^n; \hat{V}^n) \end{aligned}$$

(By Fano and data processing inequality)

$$\leq \varepsilon \log |V| + \frac{1}{n} h(\varepsilon) + \frac{1}{n} \sum_{i=1}^n I(X_i; Y_i)$$

$\left[\because X_i, Y_i - X_i - Y_i$
over DMC]

$$\leq \varepsilon \log |V| + C + \frac{1}{n} h(\varepsilon)$$

$\left[\because C \geq I(X_i; Y_i) \forall i \right]$

$$\leq_n C + \varepsilon (\log |V| + 1)$$

Thus, we have $H \leq C + \varepsilon (\log |V| + 1)$,
 for every $\varepsilon > 0$. Hence $H \leq C$.
 \square

Remark:- The achievability proof of
 the joint source-channel coding ^{the sum}
 says that source coding and channel
 coding can be optimised separately.

* Differential entropy, β -divergence, and mutual information.

Let X be a continuous random variable with the pdf f_X . Then, the differential entropy of X is defined as

$$h(X) \triangleq \int_{\text{supp}(X)} f_X(x) \log \frac{1}{f_X(x)} dx, \text{ provided the integral exists.}$$

• Lemma :- There exists X , s.t. $h(X) < 0$.

Proof:- Let $X \sim \text{unif}[a, b]$.

$$\text{Then } h(X) = \int_{[a,b]} \frac{1}{(b-a)} \log(b-a) dx$$

$$= \log(b-a).$$

Thus if $b-a < 1$, $h(X) < 0$.

Let $X_1, \dots, X_n \sim f_{X^n}$. Then, the joint differential entropy is defined as

$$h(X_1, \dots, X_n) = - \int_{\text{supp}(X_1, \dots, X_n)} f_{X^n}(x^n) \cdot \log \frac{1}{f_{X^n}(x^n)} dx^n,$$

provided the integral exists.

Let $x^{n-1} \in \text{supp}(X^{n-1})$. Then, the conditional differential entropy is defined as

$$h(X_n | X^{n-1}) = - \int_{\text{supp}(X^n)} f_{X^n}(x^n) \cdot \log \frac{1}{f_{X_n | X^{n-1}}(x_n | x^{n-1})} dx^n,$$

provided the integral exists.

• Definition:- Given an $f: [0, \infty) \rightarrow [-\infty, \infty]$,

s.t. i) f is convex

ii) $f(x) \in \mathbb{R}$ for $x \in (0, \infty)$

iii) $f(0) = 0$

iv) $\lim_{t \rightarrow 0^+} f(t) = f(0)$

Then, for any two distributions with pdfs

g_1, g_2 , with $\text{supp}(g_2) \supseteq \text{supp}(g_1)$, the f -divergence

is defined as

$$D_f(g_1 || g_2) = \int_{\text{supp}(g_2)} g_2(x) f\left(\frac{g_1(x)}{g_2(x)}\right) dx$$

Remarks:-

1. Again, by choosing $f(x) = x \log x$, we obtain the relative entropy as

$$D(g_1 || g_2) = \int_{\text{supp}(g_2)} g_1(x) \log\left(\frac{g_1(x)}{g_2(x)}\right) dx$$

2. [For more advanced students]:-

More formally, the f -divergence between

any two distributions $P_1 \ll P_2$ is given by

$$D_f(P_1 \| P_2) = \int_{\Omega} f\left(\frac{dP_1}{dP_2}\right) \cdot dP_2.$$

Note that if $P_1 \ll P_2 \ll L$ where L is the Lebesgue measure, then we have

the pdfs $g_1 = \frac{dP_1}{dL}$, $g_2 = \frac{dP_2}{dL}$, and the definition mentioned above follows.

Lemma :- $D_f(g_1 \| g_2) \geq 0$, with equality iff $g_1 = g_2$ a.e. $[g_2]$.

$$\begin{aligned}
 \text{Proof:- } D_f(g_1 \| g_2) &= \int_{\text{supp}(g_2)} g_2(x) \cdot f\left(\frac{g_1(x)}{g_2(x)}\right) \cdot dx \\
 &\geq f\left(\int_{\text{supp}(g_2)} g_2(x) \cdot \frac{g_1(x)}{g_2(x)} \cdot dx\right) \quad [\text{Jensen, since } f \text{ convex}] \\
 &= f\left(\int_{\text{supp}(g_2)} g_1(x) \cdot dx\right) \\
 &= f(D) \quad [\because \text{supp}(g_2) \supseteq \text{supp}(g_1)]
 \end{aligned}$$

$= 0$.
The inequality is satisfied with an equality

iff $\frac{g_1(x)}{g_2(x)} = c$, a.e. $[g_2]$

$$\Leftrightarrow g_1(x) = g_2(x) \text{ a.e. } [g_2]$$

C: they are pdfs and
 $\text{supp}(g_2) \supseteq \text{supp}(g_1)$

... 

Definition :- The mutual information between X and Y is defined as

$$I(X;Y) = D(f_{XY} || f_X f_Y)$$

Lemma :- $I(X;Y) \geq 0$ with equality iff

$$X \perp Y.$$

$$\begin{aligned} \text{i)} \quad I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X), \end{aligned}$$

provided the R.H.S. is well-defined

iii) $h(X) \geq h(X|Y)$ with equality
iff $X \perp Y$.

provided the quantities are well-defined.

Proof: i) follows directly from the lemma on non-negativity of f-divergence,

ii) follows from i) provided ii) is true. It remains to prove ii)

$$\begin{aligned}
 I(X;Y) &= \int_{\text{supp}(f_X \times f_Y)} f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) \cdot f_Y(y)} \cdot dx \cdot dy \\
 &= \int_{\text{supp}(f_X \times f_Y)} f_{XY}(x,y) \log f_{X|Y}(x|y) \cdot dx \cdot dy \\
 &\quad + \int_{\text{supp}(f_X \times f_Y)} f_{XY}(x,y) \log \frac{1}{f_X(x)} \cdot dx \cdot dy \\
 &= -h(X|Y) + \int_{\text{supp}(f_X)} f_X(x) \log \frac{1}{f_X(x)} \cdot dx
 \end{aligned}$$

[Here we can use the iterated integral theorem for the second integral as before, and cdfs obtained on integrating them are continuous functions and so no log. If we used the general Lebesgue integral definition, then we would require Fubini's theorem.]

$$= h(X) - h(X|Y).$$

One can similarly show $h(Y) - h(Y|X) = I(X;Y)$.
..... 

• Lemma :- i) $h(X^n) = \sum_{i=1}^n h(X_i | X^{i-1})$ [Chain rule]

ii) $h(X^n) \leq \sum_{i=1}^n h(X_i)$ [Subadditivity]

with equality iff X_1, \dots, X_n are mutually independent.

Proof: ii) would follow directly from i) and the conditioning reduces differential entropy result.

So we need to prove i).

$$\begin{aligned}
 h(X^n) &= \int_{\text{supp}(f_X)} f_{X^n}(x^n) \cdot \log \frac{1}{f_{X^n}(x^n)} \cdot dx^n \\
 &= \int_{\text{supp}(f_X)} \sum_{i=1}^n f_{X^n}(x^n) \cdot \log \frac{1}{f_{X_i|X^{i-1}}(x_i|x^{i-1})} dx^n \\
 &= \sum_{i=1}^n \int_{\text{supp}(f_{X^n})} f_{X^n}(x^n) \cdot \log \frac{1}{f_{X_i|X^{i-1}}(x_i|x^{i-1})} dx^n \\
 &= \sum_{i=1}^n \int_{\text{supp}(f_{X^i})} f_{X^i}(x^i) \cdot \log \frac{1}{f_{X_i|X^{i-1}}(x_i|x^{i-1})} dx^i
 \end{aligned}$$

[Again by iterated integral theorem

using continuity of f_{X^n} , \log and \log]

$$= \sum_{i=1}^n h(X_i | X^{i-1})$$

..... \square

The conditional mutual information between X, Y, Z is defined (as in discrete case)

as

$$I(X;Y|Z) \triangleq \mathbb{E}_Z [I(X;Y|Z=z)]$$

Remark:- One can similarly prove

- i> Non-negativity of conditional mutual information, with equality iff conditional independence.
- ii> Chain rule of mutual information
- iii> Data processing inequality.

The proofs are omitted to avoid repetition.

• Lemma :- For any X^n where $h(X^n)$ is defined,
any $c \in \mathbb{R}^n$, $h(X^n + c) = h(X^n)$

Proof :- Let $Y^n = X^n + c$.

Note that $\text{supp}(Y^n) = \text{supp}(X^n) + c$

where $A + c = \{a + c : a \in A\}$.

Now, note that

$$\begin{aligned} f_{X^n}(x^n) &= \frac{f_{Y^n}(x^n + c)}{|\det(J(\frac{x^n}{y^n})(x^n + c))|} && \forall x^n \in \text{supp}(X^n) \\ &= \frac{f_{Y^n}(x^n + c)}{|\det(J(\cdot))|} \\ &= f_{Y^n}(x^n + c) \end{aligned}$$

Now, $h(X^n + c) = h(Y^n)$

$$\begin{aligned} &= \int_{\text{supp}(Y^n)} f_{Y^n}(y^n) \log \frac{1}{f_{Y^n}(y^n)} \cdot dy^n \\ &= \int_{\text{supp}(Y^n) - c} f_{Y^n}(x^n + c) \cdot \log \frac{1}{f_{Y^n}(x^n + c)} |\det(J(\frac{y^n}{x^n})(x^n))| dx \end{aligned}$$

(By change of variables
theorem for multiple integrals)

$$= \int_{\text{supp}(x^n)} f_{x^n}(x^n) \log \frac{1}{f_{x^n}(x^n)} |\det(I)|. dx^n$$

$$\left[\because f_{x^n}(x^n) = f_{y^n}(x^n + c) \right]$$

$$\text{and } \text{supp}(x^n) = \text{supp}(y^n) - c \right]$$

$$= h(x^n).$$



• Lemma :- Let $A \in \mathbb{R}^{n \times n}$ be non-singular and X be random variable on \mathbb{R}^n . Then,

$$h(AX) = h(X) + \log |\det(A)|$$

Proof: Let $Y = AX$ and note that

A is invertible. Thus,

$$\text{supp}(Y) = A^{-1} \cdot \text{supp}(X)$$

By transformation of random variables and
noting A is invertible, we get

$$\begin{aligned}
 f_X(x) &= \frac{f_Y(Ax)}{\left| \det \left(J\left(\frac{A^{-1}y}{x}\right)(y) \right) \right|} \\
 &= \frac{f_Y(Ax)}{\left| \det(A^{-1}) \right|} \\
 &= f_Y(Ax) \cdot \left| \det(A) \right| \quad \left[\because \det(A^{-1}) = \frac{1}{\det(A)} \right]
 \end{aligned}$$

Then, $h(Ax) = h(Y)$

$$\begin{aligned}
 &= \int_{\text{supp}(Y)} f_Y(y) \cdot \log \frac{1}{f_Y(y)} \cdot dy \\
 &= \int_{A^{-1} \cdot \text{supp}(Y)} f_Y(Ax) \cdot \log \frac{1}{f_Y(Ax)} \cdot \left| \det \left(J\left(\frac{y}{x}\right)(x) \right) \right| \cdot dx
 \end{aligned}$$

[By change of variables formula]

$$= \int_{\text{supp}(X)} \frac{f_X(x)}{|\det(A)|} \log \frac{|\det(A)|}{f_X(x)} \cdot |\det(A)| \cdot dx$$

$\because \frac{f_X(x)}{|\det(A)|} = f_Y(Ax),$
 $[x \in \text{supp}(X)]$

$$= \int_{\text{supp}(X)} f_X(x) \log \frac{1}{f_X(x)} dx + \int_{\text{supp}(X)} f_X(x) \cdot \log |\det(A)| dx$$

$$= h(X) + \log |\det(A)|$$

. \square

• Lemma :- Let $X \sim N(\mu, K)$, where $K \in \mathbb{R}^{n \times n}$ is non-singular. Then,

$$h(X) = \frac{1}{2} \log [(2\pi e)^n |\det(K)|]$$

Proof :- We can assume $\mu = 0$, as otherwise

let $Y = X - \mu$. Then $Y \sim N(0, K)$

and $h(Y) = h(X)$ as differential

entropy is translation invariant.

So let $\mu=0$.

Since K is non-singular, f_x exists and

$$f_x(x) = \frac{1}{(2\pi)^{n/2} |\det(K)|^{1/2}} e^{-\frac{1}{2} x^T K^{-1} x}$$

So,

$$\begin{aligned} h(X) &= \int_{\mathbb{R}^n} f_x(u) \log \frac{1}{f_x(u)} du \\ &= \int_{\mathbb{R}^n} f_x(u) \log \left[(2\pi)^{n/2} / |\det(K)|^{1/2} \right] du \\ &\quad + \int_{\mathbb{R}^n} \left(\frac{1}{2} \log e \right) u^T K^{-1} u \cdot f_x(u) du \\ &= \frac{1}{2} \log \left[(2\pi)^n |\det(K)| \right] \\ &\quad + \frac{1}{2} \log e \mathbb{E}[X^T K^{-1} X] \end{aligned}$$

$$\text{Now, } \mathbb{E}[X^T K^{-1} X]$$

$$\begin{aligned}
&= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n X_i K_{ij}^{-1} X_j\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[X_i X_j] \cdot K_{ij}^{-1} \\
&= \sum_{i=1}^n \sum_{j=1}^n K_{ji} K_{ij}^{-1} \quad [\because \mathbb{E}[X] = 0] \\
&= \sum_{i=1}^n 1 \quad \left[\because \sum_{j=1}^n K_{ej}^{-1} K_{ji} = (K^{-1} K)_{ii} = 1\right] \\
&= n
\end{aligned}$$

Thus,

$$\begin{aligned}
h(X) &= \frac{1}{2} \log((2\pi)^n \cdot |\det(K)|^{1/2}) \\
&\quad + \frac{n}{2} \log e \\
&= \frac{1}{2} \log((2\pi e)^n \cdot |\det(K)|^{1/2})
\end{aligned}$$

∴ (3)

• Corollary:- If $X \sim N(\mu, \sigma^2)$, then

$$h(X) = \frac{1}{2} \log(2\pi e^{\sigma^2}).$$

Proof:- Follows from the previous lemma
with $n=1$, $K = [\sigma^2]$.



* Advanced (measure theoretic) definitions of
 D_f , $h(\cdot)$, and I

Note:- This section is intended to make
the notions of the 'information measures'
rigorous when the random variables
involved are not necessarily continuous.

The parts written in blue are meant for
everyone, and those in red are meant for
more advanced students.

• Definition:- Let $f: [0, \infty) \rightarrow [-\infty, \infty]$ be a function satisfying

- (i) f is convex
- (ii) $f(1) = 0$
- (iii) $f(x) \in \mathbb{R}$ if $x > 0$
- (iv) $f(0) = \lim_{t \rightarrow 0^+} f(t)$.

Then given any two measures $P \ll Q, m R,$

$$D_f(P||Q) = \int_R f\left(\frac{dP}{dQ}\right) \cdot dQ.$$

If $P \not\ll Q$, then define $\lambda = P + Q$ and note that $P \ll \lambda, Q \ll \lambda$, and define

$$D_f(P||Q) = \int_R f\left(\frac{dP/d\lambda}{dQ/d\lambda}\right) \cdot dQ.$$

Remarks:- If $P \ll Q$, then $\lambda \ll Q$.

To see this, let $Q(A) = 0$. Then

$P(A) = 0$, and hence $\lambda(A) = 0$.

Then, $\frac{dQ}{d\lambda} = \frac{1}{\frac{d\lambda}{dQ}}$, and by chain rule
of R-N derivatives $\frac{dP/d\lambda}{dQ/d\lambda} = \frac{dP}{dQ}$.

Hence, even if $P < Q$, we can define

$$D_f(P||Q) = \int_R f\left(\frac{dP/d\lambda}{dQ/d\lambda}\right) dQ$$

ii) If $P \ll Q$, i.e., there exists sets such that $P(A) > 0$, $Q(A) = 0$, then
 $D(P||Q) = \infty$.

To see this, if $P \ll Q$, then there exist A s.t. $P(A) \neq 0$, $Q(A) = 0$.

$$\text{Now, } \int_A \frac{dQ}{d\lambda} \cdot d\lambda = Q(A) = 0.$$

Now, note that $\frac{dQ}{d\lambda} \geq 0$, unless Q is a signed-measure, which of course

it is not.

Thus, we must have that $\exists B \subseteq A$ s.t.

$$\frac{dQ}{d\lambda}(\omega) = 0, \quad \forall \omega \in A \setminus B, \text{ and } Q(B) = 0.$$

Next, note that $P \ll \lambda$ and hence
 $P(B) = 0$.

Also, $\frac{dP(\omega)}{d\lambda} > 0$ for some $C \subseteq A \setminus B$,

as $P(A) > 0$. Hence $P(C) > 0$

Thus, $\frac{dP/d\lambda}{dQ/d\lambda} = \infty$, for $\omega \in C$.

$$\begin{aligned} \text{Now, } D(P||Q) &\geq \log \left(\int_{\mathbb{R}} \frac{dP/d\lambda}{dQ/d\lambda} \cdot dP \right) \text{ [Jensen's]} \\ &\geq \log \left(\int_C \frac{dP/d\lambda}{dQ/d\lambda} \cdot dP \right) \left[\because \frac{dP}{d\lambda}, \frac{dQ}{d\lambda} > 0 \right] \\ &= \infty \quad [\because P(C) > 0] \end{aligned}$$

• Definition:- Given a random variable X supported in \mathbb{R}^d ,

$$h(X) = -D(P_X || \mathcal{L}) \text{, where}$$

\mathcal{L} is the Lebesgue measure on \mathbb{R}^d .

Remark:-> The above definition ensures that if X is (jointly) continuous, then the above definition agrees with our earlier definition of $h(X)$.

ii) If X is not jointly continuous, then $h(X) = -\infty$, as per this definition

In particular, if X is discrete then

$h(X) = -\infty$. So Shannon entropy does not equate to the differential entropy.

• Definition :-

$$h(X|Y) = - \int D(P_{X|Y(y)} || \mathcal{L}) \cdot dP_Y$$

Remarks :-

1. Thus, $h(X|Y) = -\infty$ unless $P_{X|Y(y)} << \mathcal{L}$ a.e. P_Y .
2. In particular $h(X|X) = -\infty$,
 $h(f(X)|X) = -\infty$.
3. This definition reduces to the old definition when X, Y are jointly continuous.
4. If $X \perp Z$ then $h(X|Z) = h(X)$

To see this,

$$\begin{aligned} h(X|Z) &= - \int D(P_{X|Z(z)} || \mathcal{L}) \cdot dP_Z \\ &= - D(P_X || \mathcal{L}) \cdot \int dP_Z \quad [: P_{X|Z(z)} = P_X] \end{aligned}$$

$$= h(x).$$

• Definition: - Let X, Y, Z be arbitrarily jointly distributed random variables. Then

$$I(X; Y) = D(P_{XY} || P_X \times P_Y)$$

$$I(X; Y|Z) = \int D(P_{X|Y|Z}(\cdot|z) || P_{X|Z}(\cdot|z) \times P_{Y|Z}(\cdot|z)) dP_z$$

Remarks: - i) The general definition reduces to the definitions we studied for jointly continuous X, Y, Z , or discrete.

ii) $I(X; Y) \geq 0, I(X; Y|Z) \geq 0$

continues to hold for this.

To see this, note that

$$I(X; Y) = D(P_{XY} || P_X \times P_Y) \geq 0$$

$$I(X; Y|Z) = \int D(P_{X|Y|Z=z} || P_{X|Z=z} P_{Y|Z=z}) dP_z \geq 0$$

• Lemma :- $I(X; X) = H(X)$, if X is discrete
 $= \infty$, otherwise

Proof:- The discrete case we have
already shown. We need to show
it when X is non-discrete.

If X is non-discrete, define

$$A = \{x : P_x(\{x\}) > 0\}.$$

Note that A is an almost countable set as
 A corresponds to the set of discontinuities
of the cdf F_X , and F_X is a bounded
continuous function.

Note that $A = \emptyset$ is possible even if
 X is not continuous, for example if X is
the Cantor random variable.

Also, since A is at most countable,

$$A \in \mathcal{B}(\mathbb{R}).$$

Furthermore, $P_X(A) < 1$, as otherwise X is discrete.

Now, define $\Delta_A = \{(x, x) : x \in A^c\}$

$$P_{XX}(\Delta_A) = \int_{\Delta_A} dP_{XX}(x, x')$$

$$= \int_{A^c} \int_{\Delta_A(\cdot, x)} dP_{X|X}(x' | x) \cdot dP_X(x)$$

[By Product-measure theorem]

$$= \int_{A^c} P_{X|X}(\{x\} | x) dP_X(x)$$

[$\because \Delta_A(\cdot, x) = \{x\}$]

where $x \in A^c$]

$$= \int_{A^c} dP_X(x)$$

$$= P_X(A^c)$$

$$> 0 \quad (\because P_X(A) < 1)$$

On the other hand

$$\begin{aligned} P_x \times P_x(\Delta_A) &= \int_{\Delta_A} d(P_x \times P_x)(x, x') \\ &= \int_{A^c} \int_{\Delta_A(\cdot, x)} dP_x(x') \cdot dP_x(x) \\ &\quad (\text{By Product measure theorem}) \\ &= \int_{A^c} P_x(\{x\}) \cdot dP_x(x) \quad [\because \Delta_A(\cdot, x) = \{x\}] \\ &\quad \quad \quad \text{if } x \in A \\ &= 0 \quad [\because x \in A^c] \end{aligned}$$

Thus, $P_{xx} \ll P_x \times P_x$ and hence

$$I(x; x) = D(P_{xx} || P_x \times P_x) = \infty.$$

③

Lemma: Let μ, λ be finite product measures on $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2))$ where μ is the product of $\mu_1(\omega_1)$ and $\mu_2(\cdot, \omega_1)$ and λ is the product of $\lambda_1(\omega_1)$ and $\lambda_2(\cdot, \omega_1)$.

If $\mu_1 \ll \lambda_1$ and if $\mu_2(\cdot, \omega_1) \ll \lambda_2(\cdot, \omega_1)$
 a.e. $[\lambda_1]$, then $\mu \ll \lambda$ and

$$\frac{d\mu(\omega_1, \omega_2)}{d\lambda} = \frac{d\mu_1(\omega_1)}{d\lambda_1} \times \frac{d\mu_2(\omega_2, \omega_1)}{d\lambda_2}.$$

We shall refer to this lemma as the
 Radon-Nikodym derivative product (RNDP)
 lemma.

Proof:- Consider any $A \in \sigma(F_1 \times F_2)$.

$$\mu(A) = \int_{\Omega_1} \mu_2(A(\omega_1), \omega_1) \cdot d\mu_1(\omega_1)$$

$$= \int_{\Omega_1} \int_{A(\omega_1)} \frac{d\mu_2(\omega_2, \omega_1)}{d\lambda_2} \cdot d\lambda_2(\omega_2, \omega_1) \cdot \frac{d\mu_1(\omega_1)}{d\lambda_1} d\lambda_1(\omega_1)$$

(Radon - Nikodym)

$$= \int_{\Omega_1} \int_{\Omega_2} \frac{d\mu_2(\omega_2, \omega_1)}{d\lambda_2} \cdot \frac{d\mu_1(\omega_1)}{d\lambda_1} \mathbb{1}_{A(\omega_1)}$$

[the sum]

$$d\lambda_2(\omega_2, \omega_1) \cdot d\lambda_1(\omega_1)$$

$$= \int_{\Omega_1} \frac{d\mu_2(\omega_1, \omega_2)}{d\lambda} \times \frac{d\mu_1(\omega_1)}{d\lambda} 1_A \cdot d\lambda(\omega_1, \omega_2)$$

[Fubini's Theorem]

$$= \int_A \frac{d\mu_2(\omega_2, \omega_1)}{d\lambda} \cdot \frac{d\mu_1(\omega_1)}{d\lambda} \cdot d\lambda(\omega_1, \omega_2) \quad \dots \textcircled{1}$$

Thus, if $\lambda(A) = 0$, then $\mu(A) = 0$.

Hence, $\mu << \lambda$.

Thus, for any AFF,

$$\mu(A) = \int_A \frac{d\mu}{d\lambda}(\omega_1, \omega_2) d\lambda(\omega_1, \omega_2) \quad [\text{Radon-Nikodym Theorem}]$$

Hence, by ①, $\forall A \in \mathcal{F}$

$$\int_A \frac{d\mu}{d\lambda}(\omega_1, \omega_2) d\lambda(\omega_1, \omega_2) = \int_A \frac{d\mu_1}{d\lambda_1}(\omega_1) \cdot \frac{d\mu_2}{d\lambda_2}(\omega_2, \omega_1) d\lambda(\omega_1, \omega_2)$$

So,

$$\frac{d\mu}{d\lambda}(\omega_1, \omega_2) = \frac{d\mu_1}{d\lambda_1}(\omega_1) \cdot \frac{d\mu_2}{d\lambda_2}(\omega_2, \omega_1)$$

... ②

- Lemma (Chain rule & DPI) :- For any jointly distributed random variables X, Y, Z
 - $I(X;YZ) = I(X;Y) + I(X;Z|Y)$
 - If $X-Y-Z$, then $I(X;Z) \leq I(X;Y)$
with equality iff $X-Z-Y$.

Proof :- i) First, assume that $\exists A \text{ s.t. } P_{XYZ}(A) > 0$ and $P_X \times P_{YZ}(A) = 0$, i.e. $I(X;YZ) = \infty$.

We will show that at least one of $I(X;Y)$ or $I(X;Z|Y) = \infty$, and hence the result follows since mutual information is non-negative.

To proceed we define some notation.

$$A|x \triangleq \{x : \exists (y,z) \text{ s.t. } (x,y,z) \in A\}$$

$$A|x,y \triangleq \{(x,y) : \exists z \text{ s.t. } (x,y,z) \in A\}$$

$$A_x \triangleq \{(y,z) : (x,y,z) \in A\}$$

$$A_{x,y} \triangleq \{z : (x,y,z) \in A\}$$

$$A|x,y \triangleq \{x : \exists z \text{ s.t. } (x,y,z) \in A\}$$

Similar sets, such as $A|_Y, A|_Z, A|_{Y,Z}$ are all defined in the same fashion.

Now, $P_X \times P_{YZ}(A) = \int P_{YZ}(A_x) \cdot dP_X(x)$ [Product measure theorem]
 Thus, $P_X \times P_{YZ}(A) = 0$ can occur if

$$P_X(A|x) = 0 \text{ or } P_{YZ}(A_x) = 0 \text{ a.e. } [P_X]$$

If $P_X(A|x) = 0$, then

$$\begin{aligned} P_{XYZ}(A) &= \int P_{YZ|X}(A_x|x) \cdot dP_X(x) \\ &\quad [\text{Product measure theorem}] \\ &= 0 \quad [\because P_X(A|x) = 0] \end{aligned}$$

This contradicts our assumption, and
 so we must have $P_{YZ}(A_x) = 0$ a.e. $[P_X]$

Similarly, expanding

$$P_X \times P_{YZ}(A) = \int P_X(A_{yz}) dP_{YZ}$$

we can argue $P_X \times P_{YZ}(A) = 0$ & $P_{XYZ}(A) > 0$

$$\Rightarrow P_X(A_{yz}) = 0, \text{ a.e. } [P_{YZ}]$$

$$\text{Now, } P_{XYZ}(A) = \int P_{XZ|Y}(A_y|y) \cdot dP_Y(y).$$

Since $P_{XYZ}(A) > 0$, $\exists A' \subseteq A$, s.t. $A' \in \mathcal{F}$.

$$P_{XZ|Y}(A'_y|y) > 0, \forall y \in A'_Y, \text{ and } P_Y(A'_Y) > 0.$$

Note that $A'_y = A_y$ if $y \in A'_Y$, so we have

$$P_{XZ|Y}(A'_y|y) > 0, \forall y \in A'_Y.$$

Next assume, $I(X;Y) < \infty$, i.e.,

$P_{XY} \ll P_X \times P_Y$. We will show that

show that $P_{XZ|Y}(\cdot|y) \ll P_{X|Y} \times P_{Z|Y}(\cdot|y)$

$\forall y \in A'_Y$. This ensures that

$$D(P_{XZ|Y}(\cdot|y) || P_{X|Y} \times P_{Z|Y}(\cdot|y)) = \infty,$$

$\forall y \in A'_Y$, which along with the fact that $P_Y(A'_Y) > 0$ ensures that

$$I(X;Z|Y) = \infty \text{ as required.}$$

So, fix a $y \in A'_Y$.

Then, $P_{XZ|Y}(A'_y|y) > 0$ by definition of A' .

$$\text{Now, } P_{X|Y} \times P_{Z|Y} \times P_Y(A')$$

$$= \int_{A'_Y} P_{X|Y} \times P_{Z|Y}(A'_y|y) \cdot dP_Y(y)$$

[Product measure
the sum]

$$= \int_{A'_Y} \int_{A'_{Z|Y}} P_{X|Y}(A'_{y,z}|y) \cdot dP_{Z|Y}(z|y) \cdot dP_Y(y)$$

[Product measure

$$= \int_{A'_Y} P_{X|Y}(A'|_{x,y}|y) \cdot dP_Y(y)$$

theorem]

[Product measure the sum

By noting $P_{X|Y} = P_{X|YZ}$ as
the joint measure is

$$P_{X|Y} \times P_{Z|Y} \times P_Y$$

$$= P_{XY}(A'|_{XY}) \dots \textcircled{1}$$

Now, $P_X \times P_{YZ}(A')$

$$= \int_{A' \setminus YZ} P_X(A'_{y,z}) \cdot dP_{YZ}(y,z) \quad [\text{Product measure theorem}]$$

$$= \int_{A' \setminus Y} \int_{A'_{z,y}} P_X(A'_{y,z}) \cdot dP_{Z|Y}(z|y) \cdot dP_Y(y) \quad [\text{Fubini's Theorem}]$$

$$= \int_{A' \setminus Y} P_X(A'_{|X,y}) dP_Y(y) \quad [\text{Product measure theorem}]$$

$$= P_X \times P_Y(A' \setminus Y)$$

Now, by choice of A' , we must

have $P_X \times P_{YZ}(A') = 0$ as

$$P_X \times P_{YZ}(A) = 0.$$

Thus, $P_X \times P_Y(A' \setminus Y) = 0$. Then, as $I(X;Y) < \infty$

by assumption, we must have

$$P_{XY}(A'|_Y) = 0$$

Thus, by ①, $P_{X|Y} \times P_{Z|Y} \times P_Y(A') = 0$.

In other words, $\exists B \subseteq A'|_Y$ s.t.

$$\forall y \in B, P_{X|Y} \times P_{Z|Y}(A'_y|_y) = 0, P_Y(B) > 0.$$

On the other hand, we already have

that $\forall y \in B, P_{Z|Y}(A'_y|_y) > 0$.

Thus, $D(P_{Z|Y}(\cdot|_y) || P_{X|Y} \times P_{Z|Y}(\cdot|_y)) = \infty$,
 $\forall y \in B$.

Since, $P_Y(B) > 0$, we have

$$I(X; Z|Y) = \infty \text{ as required.}$$

It remains to show that

$$I(X; Y_2) = I(X; Y) + I(X; Z|Y)$$

$$\text{if } P_{XY_2} \ll P_X \times P_{Y_2}.$$

Now,

$$I(X;YZ) = \int \log \left(\frac{dP_{XYZ}}{d(P_X \times P_{YZ})} (x,y,z) \right) dP_{XYZ}(x,y,z)$$

$$= \int \log \left(\frac{dP_{XY}}{d(P_X \times P_Y)} (x,y) \right)$$

$$+ \log \left(\frac{dP_{Z|XY}}{dP_{ZY}} (x,y,z) \right) dP_{XYZ}(x,y,z)$$

(By RNDP lemma)

$$= I(X;Y) + \int \log \left(\frac{dP_{Z|XY}}{dP_{ZY}} (x,y,z) \cdot \frac{dP_{XY}}{dP_{XY}} (x,y) \right) dP_{XYZ}(x,y,z)$$

Additivity theorem and

any measure λ satisfies

$$\lambda << \lambda \text{ and } \frac{d\lambda}{d\lambda} = 1$$

$$= I(X;Y) + \int \log \left(\frac{dP_{XZ|Y}}{d(P_{XY} \times P_{ZY})} (x,y,z) \right) dP_{XYZ}(x,y,z)$$

(By RNDP lemma)

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

ii) Now assume $X - Y - Z$ holds

Then, by part (i)

$$\begin{aligned} I(X;Z) + I(X;Y|Z) &= I(X;Y) + I(X;Z|Y) \\ &= I(X;Y) \end{aligned}$$

$(\because X \perp Z | Y)$

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

$$(\because I(X;Y|Z) \geq 0)$$

Equality holds iff

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

$$\iff X - Z - Y.$$

③

Lemma :- $I(X;Y) = h(X) - h(X|Y)$ if
 $P_X \ll L$, $P_{X|Y}(\cdot|y) \ll L$ a.e. $[P_Y]$ and
 $P_{XY} \ll P_X \times P_Y$.

Proof :- Let $L \times P_Y(A) = 0$.

Then, $P_X \times P_Y(A)$

$$= \int_{A|Y} P_X(A_y) \cdot dP_Y(y) \quad [\text{Product measure}]$$

$$= \int_{A|Y} \int_{A_y} \frac{dP_X}{dL}(x) \cdot dL(x) \cdot dP_Y(y) \quad [P_X \ll L]$$

$$= \int_A \frac{dP_X}{dL}(u) \cdot d(L \times P_Y)(u, y) \quad [\text{Fubini}]$$

$$= 0 \quad [\because L \times P_Y(A) = 0]$$

Thus, $P_{XY} \ll P_X \times P_Y \ll L \times P_Y$

Again, as $P_{XY} \ll P_X \times P_Y$,

$$\begin{aligned} I(X;Y) &= \int \log \left(\frac{dP_{XY}(x,y)}{d(L \times P_Y)} \right) dP_{XY}(x,y) \\ &= \int \left\{ \log \left(\frac{dP_{XY}(x,y)}{d(L \times P_Y)} \right) - \log \left(\frac{d(P_X \times P_Y)(x,y)}{d(L \times P_Y)} \right) \right\} dP_{XY}(x,y) \end{aligned}$$

$\because P_{XY} \ll P_X \times P_Y$

$$\begin{aligned} &= \int \log \left(\frac{dP_{XY}(x,y) \times dP_Y(y)}{dL} \right) dP_{XY}(x,y) \\ &\quad - \int \log \left(\frac{dP_X(x) \times dP_Y(y)}{dL} \right) dP_{XY}(x,y) \end{aligned}$$

[By RNDP lemma]

$$= \int \log \left(\frac{dP_{XY}(x,y)}{dL} \right) dP_{XY}(x,y)$$

$$- \int \log \left(\frac{dP_X(x)}{dL} \right) dP_{XY}(x,y)$$

$\because \frac{dP_Y(y)}{dP_Y(y)} = 1$

$$= \int \int \log \left(\frac{dP_{X|Y}(x|y)}{dL} \right) \cdot dP_{X|Y}(x|y) \cdot dP_Y(y)$$

$$- \int \int \log \left(\frac{dP_X(x)}{dL} \right) \cdot dP_{Y|X}(y|x) \cdot dP_X(x)$$

(Fubini)

$$= \int D(P_{X|Y}(\cdot|y) \| L) \cdot dP_Y(y)$$

$$- \int \log \left(\frac{dP_X(x)}{dL} \right) \cdot dP_X(x)$$

$$= -h(X|Y) + h(X).$$

..... \square

Remarks :-

1. The condition that X, Y is

jointly continuous for which

we derived $I(X;Y) = h(X) - h(X|Y)$

previously, is presumed within
this case.

Q. Note here that we do not need $P_Y \ll L$ in general.

• Lemma :- For any X^n ,

$$i) h(X^n + c) = h(X^n)$$

$$ii) h(AX) = h(X) + \log |\det(A)|$$

Proof:- We only need to prove these when $P_{X^n} \ll L$, as we have already shown the proofs for the cases where $P_{X^n} \ll L$, except for singular A .

i) If $P_{X^n} \ll L$, enough to show

$$P_{Y^n} \ll L, \text{ where } Y^n = X^n + c.$$

Since $P_{X^n} \ll L$, there exists

$$B \text{ s.t. } P_{X^n}(B) > 0, L(B) = 0.$$

$$\text{Now, } P_{Y^n}(B+c) = P_{X^n}(B) > 0.$$

But $L(B+C) = L(B) = 0$ since Lebesgue measure is translation invariant.

Hence $P_{Y^n} \ll L$ as required.

ii) The following result holds for L .

For any linear transformation

$A: \mathbb{R}^n \rightarrow \mathbb{R}^n$, and any $B \in \mathcal{B}(\mathbb{R}^n)$

$$L(A \cdot B) = |\det(A)| \cdot L(B)$$

First assume A is non-singular.

So, if $P_{X^n} \ll L$, and let B be such that $L(B) = 0$, $P_{X^n}(B) > 0$.

Then, with $Y = AX$, $P_{Y^n}(B) = P_{X^n}(A^{-1}B)$

$$\text{But } L(A^{-1}B) = |\det(A)| \cdot L(B) = 0.$$

Thus, $P_{Y^n} \ll L$ and hence the result follows.

On the other hand, if A is singular, range of AX is a subspace $B \subseteq \mathbb{R}^n$ of dimension equal to $\text{rank}(A) < n$. Thus, $P_Y(B) = 1$, where $Y = AX$. But $L(B) = 0$, as $\dim(B) < n$. Thus, $h(AX) = -\infty$.

But singular $A \Rightarrow \det(A) = 0$. Thus,

$$h(AX) = h(x) + \log |\det(A)|.$$

... (2)

Lemma (Maximality of Gaussian entropy):-

Let $X \in \mathbb{R}^n$ satisfy $K = \mathbb{E}[XX^T]$ where K non-singular.

$$\text{Then, } h(X) \leq \frac{1}{2} \log \left[(2\pi e)^n |\det(K)|^{1/2} \right],$$

with equality iff $X \sim N(\mu, K)$ for some $\mu \in \mathbb{R}$.

Proof: Observe that it is enough to consider continuous X as $h(X) = -\infty$ o.w.

Also, enough to consider $\mathbb{E}[X] = 0$ using the translational invariance of $h(\cdot)$.

$\text{det } Y \sim N(0, K)$.

Then,

$$0 \leq D(f_x || f_Y)$$

$$= \int_{\mathbb{R}} f_X(x) \cdot \log \frac{f_X(x)}{f_Y(x)} \cdot dx$$

$$= -h(x) + \int_{\mathbb{R}} f_X(u) \cdot \left(\frac{1}{2} \log [(2\pi)^n \cdot |\det(K)|] \right) du$$

$$+ \frac{1}{2} \log e \int_{\mathbb{R}} x^T K^{-1} x \cdot f_X(u) du$$

$$\left[\because f_Y(y) = \frac{1}{(2\pi)^n \cdot |\det(K)|} e^{-y^T K^{-1} y} \right]$$

$$= -h(x) + \frac{1}{2} \log [(2\pi)^n \cdot |\det(K)|]$$

$$+ \frac{1}{2} \log e \mathbb{E}_x [x^T K^{-1} x]$$

$$= -h(x) + \frac{1}{2} \log [(2\pi e)^n |\det(K)|]$$

$$\left[\because \text{As in last proof } \mathbb{E}_x [x^T K^{-1} x] = n, \right]$$

since $\mathbb{E}[xx^T] = K$

$$= h(X) - h(Y)$$

$$\left[\because h(Y) = \frac{1}{2} \log \left[(2\pi e)^n |\det(K)|^{1/2} \right] \right]$$

Thus,

$h(X) \leq h(Y)$, with equality iff

$$X \sim N(\mu, K). \quad \dots \quad \text{D}$$

* Memoryless Gaussian channel

Definition:- A memoryless Gaussian channel with variance σ^2 is a system with input alphabet \mathcal{R} , output alphabet \mathcal{R} , such that for any $n \geq 1$, an input $x^n \in \mathcal{R}^n$ to this channel leads to an output $y^n \sim N(x^n, \sigma^2 I_n)$.

Remark:- In other words, this means

$$Y_i = x_i + Z_i, \text{ where } Z_1, \dots, Z_n \text{ are iid } N(0, \sigma^2).$$

• Definition :- A real number $R \geq 0$ is said to be an achievable rate for a Gaussian channel with variance σ^2 with an average power constraint P , if there exists a sequence of message sets M_n , and a sequence of encoder decoder pairs

$$f^{(n)}: M_n \rightarrow \mathbb{R}^n \text{ and } g^{(n)}: \mathbb{R}^n \rightarrow M_n \text{ s.t.}$$

$$(i) \liminf_{n \rightarrow \infty} \frac{\log |M_n|}{n} = R$$

$$(ii) \lim_{n \rightarrow \infty} P_m(W \neq g^{(n)}(f^{(n)}(w))) = 0, \forall w \in M_n$$

$$(iii) \frac{1}{n} \sum_{i=1}^n |f^{(n)}(w)(i)|^2 \leq P, \quad \forall w \in M_n, \quad \forall n \geq 1.$$

Remark :- 1. The constraint number (ii) can be replaced by

$$\lim_{n \rightarrow \infty} \max_{w \in M_n} P_n(w \neq f^{(n)}(f^{(n)}(w))) = 0,$$

just as in the case of DMC. This will not affect the capacity.

2. The constraint (iii) is a peculiarity of the Gaussian or any other channel with unbounded alphabets.

Note that (iii) is an average power constraint. Here, unlike (ii) we cannot replace it with a max power constraint of the form

$$f^{(n)}(w)(i)^2 \leq P, \quad \forall i \in [n], w \in M_n, n \geq 1.$$

If we do say, then the expression for capacity changes, and the capacity achieving distribution is in fact discrete.

• Definition:- The capacity of a memoryless Gaussian channel with variance σ^2 and average power constraint P , denoted by $C(P, \sigma^2)$ is defined as the supremum of all achievable rates.

• Gaussian Channel coding theorem :-

$$C(P, \sigma^2) = \sup_{\substack{P_X: E[X^2] \leq P, \\ X \perp Z}} I(X; X+Z) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right),$$

where $Z \sim N(0, \sigma^2)$.

We shall first build some machinery to prove this theorem.

• Lemma :- Let $X \perp Z$, and let Z be a continuous random variable. Then,

$$i) P_{X+Z, X} \ll P_{X+Z} \times P_X$$

$$ii) P_{X+Z} \ll L$$

$$iii) P_{X+Z|X}(\cdot|x) \ll L$$

$$iv) I(X; X+Z) = h(X+Z) - h(Z).$$

For advanced students only.

Proof:

[For advanced students]

Firstly note that for any set B ,

$$P_{X+Z|X}(B|x) = P_{Z|X}(B-x|x)$$

$$\left[\text{where } B-x = \{x' : x+x' \in B\} \right]$$

$$= P_Z(B-x) \quad (\because Z \perp X)$$

i) Let A be any set such that

$$P_{X+Z} \times P_X(A) > 0$$

Then, $\int P_{X+Z}(A_x) \cdot dP_X(x) = 0$ [where $A_x = \{y : (y, x) \in A\}$]

$$\Rightarrow \int P_Z(A_x - x) \cdot dP_X(x) = 0 \quad (\text{Product measure})$$

$$\Rightarrow \int P_{z+x|x} (A_x|x) \cdot dP_x(x) = 0$$

$$\Rightarrow P_{x+z,x}(A) = 0$$

Thus $P_{x+z,x} \ll P_{x+z} \times P_x$

ii) Let A be any set such that $L(A)=0$.

Then, $P_{x+z}(A)$

$$= P_z(A-x) \quad [: X \perp Z]$$

Now, $L(A-x)=0$ as $L(A)=0$.

Then, $P_z \ll L \Rightarrow P_z(A-x)=0$.

Thus, $P_{x+z}(A)=0$.

So, $P_{x+z} \ll L$.

iii) Again, let $L(A)=\beta$.

$$P_{x+z|x}(A|x) = P_z(A-x)$$

Again as $L(A-x)=0$ since $L(A)=0$

$$P_z(A-x)=0 \text{ as } P_z \ll L.$$

Thus $P_{x+z|x}(A|x) = 0$.

Hence, $P_{x+z|x}(\cdot|x) \ll L$.

iv) By parts (i)-(iii), we can write

$$I(X;X+z) = h(X+z) - h(X+z|X)$$

Now, $h(X+z|x)$

$$= - \int D(P_{x+z|x}(\cdot|x) \| L) \cdot dP_x(x)$$

$$= - \int D(P_{z'_x} \| L) \cdot dP_x(x)$$

[where $z'_x = Z+x$ since

$$P_{x+z|x}(A|x) = P_z(A-x)$$

for all sets]

$$= \int h(z) \cdot dP_x(x) \quad [\because h(Z+x) = h(z)]$$

$$= h(z).$$

$$\Rightarrow I(X;X+z) = h(X+z) - h(z).$$

. . . \square

Lemma :- Let Z_1, \dots, Z_n are i.i.d and $Z^n \perp X^n$. Then if Z_1, \dots, Z_n are jointly continuous,

$$\text{i} \triangleright P_{X^n+Z^n} \ll L \quad [\text{i.e. } X^n+Z^n \text{ has a pdf}]$$

$$\text{ii} \triangleright P_{X^n+Z^n|X^n}(\cdot|x^n) \ll L \quad [\text{i.e. } f_{X^n+Z^n|X^n}(\cdot|x) \text{ exists}]$$

$$\text{iii} \triangleright P_{X^n+Z^n, X^n} \ll P_{X^n+Z^n} \times P_{X^n}$$

$$\text{iv} \triangleright I(X^n; X^n+Z^n) = h(X^n+Z^n) - n h(Z)$$

where $X^n+Z^n \triangleq (Y_1, \dots, Y_n)$ where

$Y_i = X_i + Z_i$, and Z is a random variable with same distribution as $Z_{i,n}$.

Proof:- The proof of i), ii), iii) are exactly the same as the previous lemma using the key step that

$$P_{X^n+Z^n|X^n}(A_{X^n}|x^n) = P_{Z^n}(A_{X^n}-x^n)$$

$$= P_{X^n + Z^n}(A_{x^n}),$$

for any measurable A , and any x^n ,
where $A_{x^n} = \{z^n : (x^n + z^n, x^n) \in A\}$,
where $x^n + z^n \triangleq (x_i + z_i : i \in [n])$.

Thus, by a previous lemma, we can

expand $I(x^n; X^n + Z^n) = h(X^n + Z^n) - h(X^n + Z^n | X^n)$

It remains to show $h(X^n + Z^n | X^n) = nh(Z)$.

To see this,

$$h(X^n + Z^n | X^n) = -\int D(P_{X^n + Z^n | X^n}(\cdot | x^n) \| \mathcal{L}) dP_X(x^n)$$

But, note that for any x^n and any A ,

$$P_{X^n + Z^n | X^n}(A | x^n) = P_{Z^n}(A - x^n) = P_{Z'^n}(A),$$

$$\text{where } Z'^n = Z^n + x^n$$

Since translation doesn't alter differential entropy, we have $D(P_{X^n + Z^n | X^n}(\cdot | x^n) \| \mathcal{L})$
 $= D(P_{Z'^n} \| \mathcal{L}) = -h(Z^n) = -nh(Z)$,

$\because Z^n$ are i.i.d and continuous.]

$$\text{So, } h(X^n + Z^n | \cancel{X}^n) = nh(Z). \quad \dots \square$$

Lemma :- $\sup_{P_X} I(X; X+Z) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right)$

$P_X: \mathbb{E}[X^2] \leq P$

where $Z \sim N(0, \sigma^2)$, $Z \perp X$.

Proof :- Since $Z \perp X$ and
 Z is continuous we have by
previous lemma that

$$I(X; X+Z) = h(X+Z) - h(Z)$$

$$\leq \frac{1}{2} \log [2\pi e (P + \sigma^2)]$$

$$- \frac{1}{2} \log [2\pi e \sigma^2]$$

$$\because \mathbb{E}[(X+Z)^2]$$

$$= \mathbb{E}[X^2] + \mathbb{E}[Z^2] + 0$$

as $X \perp Z$ and
and $\mathbb{E}[X^2] \leq P$, and $\mathbb{E}[Z] = 0$,

using the maximality of
Gaussian differential entropy]

$$\text{So, } I(X; X+Z) \leq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right)$$

Note that we have an equality if and only if $X \sim N(0, P)$.

$$\text{Thus, } \sup_{P_X : E[X^2] \leq P} I(X; X+Z) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) \quad \dots \quad \square$$

* Converse of the Gaussian channel coding theorem

We begin by proving the following lemma.

Lemma: - Let X^n be an arbitrary joint distributed random vector and $Y_i = X_i + Z_i$, $\forall i \in [n]$, where Z_i are iid continuous random variables with $Z^n \perp X^n$.

Then,

$$I(X^n; Y^n) \leq \sum_{i=1}^n I(X_i; Y_i).$$

Proof:- As $X^n \perp Z^n$ and

Z^n is continuous, we have by
a previous lemma that-

$$I(X^n; Y^n) = I(X^n; X^n + Z^n)$$
$$= h(X^n + Z^n) - n \cdot h(Z),$$

where Z has

some distributions

as Z_i .

Also, the same lemma says that-

$Y^n = X^n + Z^n$ are jointly continuous.

Then by subadditivity of differential
entropy, we have

$$I(X^n; Y^n) \leq \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i)$$

$\left[\because nh(Z) = \sum_{i=1}^n h(Z_i) \right]$

$$\begin{aligned}
 &= \sum_{i=1}^n (h(Y_i) - h(Z_i)) \\
 &= \sum_{i=1}^n I(X_i; X_i + Z_i) \\
 &\quad [\because Z_i \text{ are continuous} \\
 &\quad \text{and } X_i \perp Z_i] \\
 &= \sum_{i=1}^n I(X_i; Y_i)
 \end{aligned}$$

... ④

Fix any $\varepsilon > 0$. By Fano's inequality, we have

$$\begin{aligned}
 H(W|\tilde{W}) &\leq P_e \log |M_n| + h(P_e) \\
 &\leq \varepsilon \cdot \log |M_n| + h(\varepsilon) \\
 &\quad [\because P_e \leq \varepsilon]
 \end{aligned}$$

... ①

Now,

$$H(w|\hat{w}) = \log |M_n| - I(w; \hat{w}) \\ \geq \log |M_n| - I(x^n; \hat{y}^n)$$

$$\geq \log |M_n| - \sum_{i=1}^n I(x_i; y_i) \quad \dots \quad (2)$$

[$\because X_{+i}, Y_{+i} = X_i - Y_i$ and
properties of mutual
info such as chain rule
carry over to the definition
with arbitrary random
variables]

Now, by the power constraint, for any $w \in M_n$

$$\sum_{i=1}^n f^{(n)}(w)(i)^2 \leq nP.$$

Now, $X_i = f^{(n)}(w)(i)$ and hence

$$\sum_{i=1}^n X_i^2 \leq nP \Rightarrow \sum_{i=1}^n \mathbb{E}[X_i^2] \leq nP.$$

let $\mathbb{E}[x_i^2] = P_i$.

Then, in ②, we have

$$H(w/\hat{w}) \geq \log |M_n| - \sum_{i=1}^n \frac{1}{2} \log \left(1 + \frac{P_i}{\sigma^2} \right)$$

$$\left[\because I(x_i; Y_i) \leq \frac{1}{2} \log \left(1 + \frac{P_i}{\sigma^2} \right) \right]$$

$$\text{as } \mathbb{E}[x_i^2] = P$$

$$= \log |M_n| - \frac{n}{2} \cdot \sum_{i=1}^n \frac{1}{n} \cdot \log \left(1 + \frac{P_i}{\sigma^2} \right)$$

$$\geq \log |M_n| - \frac{n}{2} \log \left[\sum_{i=1}^n \frac{1}{n} \left(1 + \frac{P_i}{\sigma^2} \right) \right]$$

[Jensen's inequality
since \log is concave]

$$\geq \log |M_n| - \frac{n}{2} \log \left(1 + \frac{P}{\sigma^2} \right)$$

[$\because \sum_{i=1}^n P_i \leq nP$ and
 \log is increasing]

$$= \log |M_n| - n \cdot C(P, \sigma^2)$$

Plugging this back in ①, we get

$$\frac{1}{n} \log |M_n| - C(P, \sigma^2) \leq \varepsilon \cdot \frac{\log |M_n|}{n} + \frac{h(\varepsilon)}{n}$$

$$\Rightarrow \frac{\log |M_n| / (1-\varepsilon)}{n} \leq C(P, \sigma^2) + \frac{h(\varepsilon)}{n}$$

Taking \liminf , we get

$$R(1-\varepsilon) \leq C(P, \sigma^2).$$

Since, choice of $\varepsilon > 0$ was arbitrary,

we must have $R \leq C(P, \sigma^2)$.

* Achievability Proof of the Gaussian channel coding theorem

We first begin by defining typical

sets. Let X, Y be jointly continuous r.v.s.

$$A_{\varepsilon, XY}^{(n)} \triangleq \left\{ (x^n, y^n) : \begin{array}{l} \left| \frac{1}{n} \log \frac{1}{f_{XY}(x^n, y^n)} - h(X, Y) \right| \leq \varepsilon, \\ \left| \frac{1}{n} \log \frac{1}{f_X(x^n)} - h(X) \right| \leq \varepsilon, \\ \left| \frac{1}{n} \log \frac{1}{f_Y(y^n)} - h(Y) \right| \leq \varepsilon \end{array} \right\}$$

Lemma (AEP for Gaussians):-

$$(i) P_n(A_{\varepsilon, XY}^{(n)}) \rightarrow 1 \text{ as } n \rightarrow \infty$$

(ii) If $\tilde{X}^n, \tilde{Y}^n \sim \text{i.i.d } P_X \times P_Y$ then

$$P_n(\tilde{X}^n, \tilde{Y}^n \in A_{\varepsilon, XY}^{(n)}) \leq 2^{-n(I(X; Y) - 3\varepsilon)}$$

Proof:- (i)

$$\frac{1}{n} \log \frac{1}{f_{XY}(x^n, y^n)}$$

$$= \frac{1}{n} \sum_{i=1}^n \log \frac{1}{f_{XY}(x_i, y_i)} \xrightarrow{P_n} \mathbb{E} \left[\log \frac{1}{f_{XY}(X, Y)} \right] = h(X, Y)$$

[Weak Law of Large numbers]

Thus, $\Pr(A_{\varepsilon, XY}^{(n)}) \rightarrow 1$ as $n \rightarrow \infty$.

$$(ii) \quad \Pr(\tilde{X}^n, \tilde{Y}^n \in A_{\varepsilon, XY}^{(n)}) \\ = \int_{A_{\varepsilon, XY}^{(n)}} f_X(x^n) \cdot f_Y(y^n) \cdot dx^n dy^n$$

[Note that this step requires

$A_{\varepsilon, XY}^{(n)}$ to be a measurable set.

This is true since by Radon-Nikodym theorem all pdfs are measurable functions, and since $\log \frac{1}{f_X}$ is also measurable, $\log \frac{1}{f_X(x^n)}$, $\log \frac{1}{f_Y(y^n)}$, $\log \frac{1}{f_{XY}(x^n, y^n)}$ are all measurable maps.

The measurability of $A_{XY, \varepsilon}^{(n)}$ now follows by definition of measurable maps]

$$\leq \int_{A_{\varepsilon,xy}^{(n)}} 2^{-n(h(x)-\varepsilon)} 2^{-n(h(y)-\varepsilon)} \cdot dx^n dy^n$$

[By defⁿ of $A_{\varepsilon,xy}^{(n)}$]

$$= 2^{-n(h(x)+h(y)-2\varepsilon)} \cdot \int_{A_{\varepsilon,xy}^{(n)}} dx^n dy^n$$

..... ①

On the other hand,

$$\Pr(A_{\varepsilon,xi}^{(n)}) \leq 1$$

$$\Rightarrow \int_{A_{\varepsilon,xi}^{(n)}} f_{xy}(x^n, y^n) \cdot dx^n dy^n \leq 1$$

$$\Rightarrow \int_{A_{\varepsilon,xy}^{(n)}} 2^{-n(h(x,y)+\varepsilon)} dx^n dy^n \leq 1$$

[By definition of
 $A_{\varepsilon,xy}^{(n)}$]

$$\Rightarrow \int_{A_{\varepsilon,XY}^{(n)}} d^n x d^n y \leq 2^{n(h(X,Y) + \varepsilon)}$$

Plugging this back in ① we get

$$\begin{aligned} & \Pr(\tilde{X}^n, \tilde{Y}^n \in A_{\varepsilon,XY}^{(n)}) \\ & \leq \int_{A_{\varepsilon,XY}^{(n)}} 2^{-n(h(X) + h(Y) - h(X,Y) - 3\varepsilon)} \\ & = 2^{-n(I(X;Y) - 3\varepsilon)} \end{aligned}$$

... \square

Remark: - 1. Notice that this entire proof heavily relies on the fact that X, Y are jointly continuous.

We now describe the coding scheme that achieves capacity.

Fix any $\epsilon > 0$ and a $\delta > 0$, and any

$$R \geq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) - 3\delta - f(\delta), \text{ where } f(\delta) = o(\delta) \text{ is specified later.}$$

We will construct a sequence of encoder-decoder pairs $f^{(n)}: M_n \rightarrow R^n$, $g^{(n)}: R^n \rightarrow M_n$, satisfying

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log |M_n| = R,$$

$$P_n(\hat{W} \neq W) \leq 2\epsilon \text{ and } \frac{1}{n} \sum_{i=1}^n f^{(n)}(w)(i)^2 \leq \frac{P}{n}, \forall w \in M_n,$$

$$\text{where } \hat{W} = g^{(n)}(f^{(n)}(w)).$$

Since choice of ϵ was arbitrary, this would prove $\liminf_{n \rightarrow \infty} \frac{\log |M_n|}{n} = R$ is an achievable rate.

Finally, the choice of $\delta > 0$ was arbitrary and hence any $R \leq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right)$ is achievable.

Thus, $C(P, \sigma^2) \leq \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right)$ which will complete the proof of the achievability part.

We begin by creating a distribution over the codebooks.

The codebooks $\mathcal{L} = \{X^{(1)}, \dots, X^{(M_n)}\}$

is distributed according to

$$N(0, (P-\delta) I_n) \times \dots \times N(0, (P-\delta) I_n).$$

Define $f^{(n)}(w) = X^n(w)$

The decoder is defined as follows:

On receiving y^n decode to

- w if $(x^n(w), y^n) \in A_{\delta, XY}^{(n)}$ and no other

w' satisfies $(x^n(w'), y^n) \in A_{\delta, XY}^{(n)}$

- \emptyset o.w.

[Note that since X^n is Gaussian, $A_{\delta, XY}^{(n)}$ is well-defined]

Define the error event \mathcal{E} as

$$\mathcal{E} = \{\hat{w} \neq w\} \cup \left\{ \frac{1}{n} \sum_{i=1}^n X_i(w) > p \right\}$$

Note that we have used an "averaged" power constraint in the error term here, and this is not enough, as this means the power constraint is averaged not only about n but also N . We shall take care of this later.

$$P_n(\mathcal{E}) = \frac{1}{|M_n|} \sum_{w \in M_n} P_r(\mathcal{E}|W=w)$$

Define $\mathcal{E}_w = \{(X^n(w), Y^n) \in A_{\delta, XY}^{(n)}\}$

Then

$$\mathcal{E}|W=w = \bigcup_{w' \neq w} \mathcal{E}_{w'} \cup \mathcal{E}_w^c \cup \left\{ \frac{1}{n} \sum_{i=1}^n X_i(w) > p \right\}.$$

Now, $X_1(\omega), \dots, X_n(\omega) \sim i.i.d N(0, P-\delta)$

Thus, by WLLN,

$$P_n \left(\left| \frac{1}{n} \sum_{i=1}^n X_i^2(\omega) - P + \delta \right| > \delta \right) \leq \frac{\epsilon}{3}$$

$$\Rightarrow P_n \left(\frac{1}{n} \sum_{i=1}^n X_i^2(\omega) > P \right) \leq \frac{\epsilon}{3} \dots \textcircled{1}$$

By the AEP for Gaussians,

$$P_n(\mathcal{E}_\omega) \leq \frac{\epsilon}{3} \dots \textcircled{2}$$

Finally,

$$P_n \left(\bigcup_{w' \neq w} \mathcal{E}_{\omega'} \mid W=w \right)$$

$$\leq \sum_{w' \neq w} P_n(\mathcal{E}_{\omega'} \mid W=w)$$

$$\leq \sum_{w' \neq w} 2^{-n(I(X; X+Z) - 3\delta)}$$

$\because \forall W=w,$

$X^n(\omega') \perp Y^n,$

and hence by AEP
for Gaussians)

$$\leq 2^{-n(I(X; X+Z) - 3\delta - R)}$$

$$= 2^{-n\left(\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) - o(\delta) - 3\delta - R\right)}$$

$$\left(\because I(X; X+Z) = \frac{1}{2}\log\left(1 + \frac{P-\delta}{\sigma^2}\right)\right)$$

$$= \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) - o(\delta)$$

$$\leq_n \frac{\epsilon}{3} \quad \left[\because R \leq \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) - 3\delta - o(\delta)\right]$$

Thus, we have from ①, ②, and ③

$$P_n(\varepsilon | W=w) \leq_n \epsilon.$$

$$\text{Hence, } P_n(\varepsilon) = \frac{1}{|M_n|} \sum_{i=1}^{|M_n|} P_n(\varepsilon | W=w) \quad \left(\because W \text{ is uniform }\{M_n\}\right)$$

$$\leq_n \epsilon$$

$$\text{Now, } P_n(\varepsilon) = E_e [P(\varepsilon | e)]$$

Thus, there exists a sequence of codebooks

ℓ_n satisfying

$$P_n(\varepsilon | \ell_n) \leq_n \varepsilon.$$

Note that this doesn't complete the proof since ℓ_n only ensures that

$$P_n\left(\frac{1}{n} \sum_{i=1}^n f^{(n)}(w)(i)^2 > P\right) \leq_n \varepsilon.$$

But not $\max_{w \in M_n} \frac{1}{n} \sum_{i=1}^n f^n(w)(i)^2 \leq P$.

Now, rank the w in M_n in increasing order of $P_n(\varepsilon | w=w, \ell_n)$, and WLOG assume $P_n(\varepsilon | w=1, \ell_n) \leq \dots \leq P_n(\varepsilon | w=|M_n|, \ell_n)$, and define

message set $M_n' = \{1, 2, \dots, \lceil \frac{1}{2} [2^{nR}] \rceil\}$,

and a new codebook ℓ_n' for M_n' which uses the same encodings as ℓ_n .

Now, it is easy to see that $\lim_{n \rightarrow \infty} \frac{1}{n} \log |M_n'| = R$

Also,

$$\varepsilon \geq_n \Pr(\varepsilon | e_n)$$

$$= \frac{1}{|M_n|} \sum_{w \in M_n} \Pr(\varepsilon | e_n, W=w)$$

$$\geq \frac{1}{|M_n|} \sum_{w \in M_n'} \max_{w' \in M_n'} \Pr(\varepsilon | e_n, W=w)$$

[By definition of M_n']

$$\geq \frac{1}{|M_n|} \sum_{w \in M_n'} \max_{w' \in M_n'} \Pr(\varepsilon | e_n', W=w)$$

[$w' \sim \text{unif}\{M_n'\}$]

$$= \frac{|M_n'|}{|M_n|} \cdot \max_{w \in M_n'} \Pr(\varepsilon | e_n', W=w)$$

$$\geq \frac{1}{2} \max_{w \in M_n'} \Pr(\varepsilon | e_n', W=w)$$

$$\Rightarrow \max_{w \in M_n'} P_n(\mathcal{E} | \ell_n', w' = w) \leq_n 2\epsilon$$

Now, Recall that $\mathcal{E} = \{\hat{w}' \neq \hat{w}\} \cup \left\{ \frac{1}{n} \sum_{i=1}^n x_i^2(w') > p \right\}$

So, given ℓ_n' and $w' = w$, the event $\left\{ \frac{1}{n} \sum_{i=1}^n x_i^2(w') > p \right\}$ is trivial (i.e., has either probability 0 or 1).

Since $\max_{w \in M_n'} P_n(\mathcal{E} | \ell_n', w' = w) \leq_n 2\epsilon$,

thus we must have

$$\frac{1}{n} \sum_{i=1}^n f^{(n)}(w)(i)^2 > p \text{ for all } w \in M_n'.$$

$$\begin{aligned} \text{Also, we have } & \max_{w \in M_n'} P_n(\mathcal{E} | \ell_n', w' = w) \\ & \geq \max_{w \in M_n'} P_n(w' \neq \hat{w} | \ell_n', w' = w) \end{aligned}$$

and so we have

$$\max_{w \in M_n'} P_n(w' \neq \hat{w} | \ell_n', w' = w) \leq_n \epsilon.$$

This completes the proof.

Remarks:- Note that we needed to actually prove the scenario with max probability error in order to satisfy the power constraint.