▸ Jensen's Inequality

A function $f(x)$ will be called a convex function over an interval (a,b) if for $x_1, x_2 \in (a,b)$ and $0 \leq \lambda < 1$

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

Then $-f$ is concave.

Thm: If the functions 2nd derivative is ≥ 0 everywhere, then we call function to be convex.

Proof: $f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0^*)}{2}(x-x_0)^2$ → Taylor exp. around x_0

$x_0 < x^* < x$

$\Rightarrow \; f(x) - f(x_0) - f'(x_0)\cdot(x-x_0)$.

$f(x) \geq f(x_0) + f'(x_0)(x-x_0)$

1st case

take $x = x_1$, $x_0 = \lambda x_1 + (1-\lambda)x_2$

$f(x_1) \geq f(x_0) + f'(x_0)[x_1 - \lambda x_1 - (1-\lambda)x_2]$.

$x_1 - \lambda x_1 - x_2 + \lambda x_2$

$= (x_1 - x_2) - \lambda(x_1 - x_2) = (1-\lambda)(x_1 - x_2)$.

$* \; f(x_1) \geq f(x_0) + f'(x_0)[(1-\lambda)(x_1 - x_2)]$ —— (I)

2nd case:

$x = x_2$, $x_0 = \lambda x_1 + (1-\lambda)x_2$ → $f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]$. —— (II).

multiplying (I) by λ & (II) by $(1-\lambda)$ & then adding

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq \lambda f(x_0) + \cancel{\lambda(1-\lambda) f'(x_0)(x_1 - x_2)}$$
$$+ (1-\lambda) f(x_0) + \cancel{\lambda(1-\lambda) f'(x_0)(x_2 - x_1)}$$

$\Rightarrow \; \lambda f(x_1) + (1-\lambda) f(x_2) \geq \cancel{\lambda f} \; f(x_0)$.

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

$\Rightarrow \; \boxed{\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)}$

→ f is convex.

If we have a convex function f and X is a random variable, then $\boxed{E[f(X)] \geq f(E[X])}$ → Jensens Inequality.

Proof for 2 mass point distribution, it directly follows from being convex.
$[p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)]$

Let this be true for k-1 mass point. & let $p_i' = \frac{p_i}{1-p_k}$

$$\Rightarrow \sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k) \cdot \left\{ \sum_{i=1}^{k-1} p_i' f(x_i) \right\}$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right).$$

$$\geq f\left(p_k x_k + (1-p_k) \cdot \sum_{i=1}^{k-1} p_i' x_i\right).$$

$$\geq f\left(\sum_{i=1}^{k} p_i x_i\right).$$

Note

$\sum p_i' =$

$\frac{p_1 + p_2 \cdots p_{k-1}}{1-p_k} = \frac{1-p_k}{1-p_k}$

$= 1$

Now since

$$\sum_{i=1}^{k} p_i f(x_i) \geq f\left(\sum_{i=1}^{k} p_i x_i\right)$$

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

since $\boxed{\sum_{i} p_i = 1}$

Q. Show that $H(X) \leq \log|\mathcal{X}|$

where $|\mathcal{X}|$ denotes no. of elements in the range of X

with equality iff X has uniform distribution.

Let X have uniform dist:

$H(X) = -\left(\frac{1}{n}\log\left(\frac{1}{n}\right) + \frac{1}{n}\log\left(\frac{1}{n}\right) + \cdots \frac{1}{n}\log\left(\frac{1}{n}\right)\right)$

$p_1 \quad p_1 \quad p_2$

at $\left.\begin{matrix} p_1 \to p_1 + \epsilon \\ p_2 \to p_1 - \epsilon \end{matrix}\right\} \Rightarrow -H'(X) = -\left(\left(\frac{1}{n}+\epsilon\right)\log\left(\frac{1}{n}+\epsilon\right) + \left(\frac{1}{n}-\epsilon\right)\log\left(\frac{1}{n}-\epsilon\right) + \cdots\right)$

$\sum p(x) \log n(p(x))$

$= p(x)\log n + p(x)\log p(x)$

$\log n + -H(X)$

Sol^n Let $u(x)$ be uniform dist $u(x) = \frac{1}{n}\ \forall\ (x \in X)$ $(n = |X|)$

$p(x)$ be any prob. dist of X.

$$\Rightarrow D(p\|u) = \sum_{x \in X} p(x).\log\frac{(p(x))}{(u(x))} = \sum p(x).\log(n.p(x))$$

$$= \sum p(x).\log n + \sum p(x)\log(p(x))$$

$$D(p\|u) = \log n - H(x) \geq 0$$

$$\Rightarrow \boxed{H(x) \leq \log n}$$

▶ Independence Bound on Entropy

Let $X_1, X_2, \dots X_n$ be according to prob. dist. $p(x_1, x_2, \dots x_n)$.

Then $\exists$

$$H(X_1, X_2, \dots X_n) \leq \sum_{i=1}^{n} H(X_i) \quad \Big\}\ \text{Equality holds if } X_i \text{ are independent}$$

Proof: $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$

$\cancel{H(X_1, X_2, \dots X_n) = H(X_1) + H(X_2) + H(}$

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

and $H(X_2|X_1) = H(X_1, X_2) - H(X_1)$.

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

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

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

▶ Log Sum Inequality

For non-negative numbers $a_1, a_2, \dots a_n$ and $b_1, b_2, \dots b_n$

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

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

$$\sum_{i=1}^{n}\frac{a_i}{\Sigma a_i}$$

Proof: without loss of generality, we may assume $a_i \geq 0$ and $b_i > 0$

Let us consider the function $f(x) = x\cdot\log(x)$

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

$$f''(x) = \frac{\log e}{x} > 0 \quad \forall\ x > 0$$

Hence by Jensen's inequality.

$$\Sigma \alpha_i f(x_i) \geq f(\Sigma \alpha_i x_i)$$

for $\alpha_i \geq 0$, ~~$\Sigma\alpha_i = 1$~~ Let $\alpha_i = \frac{b_i}{\sum_j b_j}$, $x_i = \frac{a_i}{b_i}$

$$\Rightarrow \sum_i \frac{b_i}{\sum_j b_i}\frac{a_i}{b_i}\log\frac{a_i}{b_i} \geq f\left(\Sigma \frac{\cancel{b_i}}{\sum_j b_j}\frac{a_i}{\cancel{b_i}}\right)$$

$$\Rightarrow \sum_i \frac{a_i}{\cancel{\sum_j b_j}}\log\frac{a_i}{b_i} \geq \sum_i \frac{a_i}{\cancel{\sum_j b_j}}\log\frac{\Sigma a_i}{\Sigma b_i}$$ HP

$\Rightarrow$ Now let $\boxed{a_i/b_i = k}$ $\forall\ i$

$$\Sigma a_i \log\left(\frac{k\cdot \cancel{b_i}}{\cancel{b_i}}\right) \geq (\Sigma a_i).\log\frac{k\cancel{\Sigma b_i}}{\cancel{\Sigma b_i}}$$

$\Rightarrow$ equality holds.

Prove $D(p\|q) \geq 0$ from this case

Let $p(x) = a_i$ & $q(x) = b_i$ s.t. $\Sigma p(x) = \Sigma q(x) = 1$

then using log sum inequality

we can directly get

$$D(p\|q) \geq 1.\log(1/1) \geq 0$$

Theorem: The relative entropy $D(p||q)$ is convex for the pair (p, q) if
$p \to p_1, p_2$, $q \to q_1, q_2$

$$\boxed{D(\lambda p_1 + (1-\lambda) p_2 \,||\, \lambda q_1 + (1-\lambda) q_2) \leq \lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2)}$$

$0 < \lambda < 1$

$a_1 = \lambda p_1(x)$ $\qquad b_1 = \lambda q_1(x)$

$a_2 = (1-\lambda) p_2(x)$ $\qquad b_2 = (1-\lambda) q_2(x)$.

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1+a_2) \cdot \log \frac{a_1+a_2}{b_1+b_2}$$

$$\rightarrow \sum \lambda p_1(x) \log \frac{\cancel{\lambda} p_1(x)}{\cancel{\lambda}\, q_1(x)} + (1-\lambda) p_2(x) \cdot \log \frac{\cancel{(1-\lambda)} p_2(x)}{\cancel{(1-\lambda)} q_2(x)} \geq$$

$$\geq \sum (\lambda p_1 + (1-\lambda) p_2) \log \frac{\lambda p_1 + (1-\lambda) p_2}{\lambda q_1 + (1-\lambda) q_2}$$

taking Σ on both sides

$$\lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2) \geq \underbrace{\sum (\lambda p_1 + (1-\lambda) p_2) \log \frac{(\lambda p_1 + (1-\lambda) p_2)}{(\lambda q_1 + (1-\lambda) q_2)}}_{D(\lambda p_1(x) + (1-\lambda) p_2(x) \,||\, \lambda q_1(x) + (1-\lambda) q_2(x))}.$$

$$\Rightarrow D(\lambda p_1 + (1-\lambda) p_2 \,||\, \lambda q_1 + (1-\lambda) q_2) \leq \lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2)$$

Hence Proved

Show that: $H(p)$ is a concave function of p

Let $f = -H(p) \longrightarrow$ ~~prove~~ convex $\quad f(p) = \Sigma p(x) \log p(x)$.

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

$p(x) \mid u(x) \rightarrow \frac{1}{|X|}$

$D(p||q) = \log(|X|) - H(p)$

$H(p) = \log |X| - D(p||q)$.

$\hookrightarrow$ concave function.

(*)

▸ The Uniqueness Theorem

1) For a given n and a prob. distribution P, $\left(\sum_{i=1}^{n} p_i = 1\right)$, the function $H(p_1, p_2, \dots p_n)$ takes the highest value at $p_k = \frac{1}{n} \ \forall\, k \in \{1, 2, \dots n\}$.

2.) $H(AB) = H(A) + H(B)$

3). $H(p_1, p_2, \dots p_n, 0) = H(p_1, p_2, \dots p_n)$

[If we add an impossible event (or events) to a prob. dist (scheme), it does not change entropy].

<u>Thm</u>: Let $H(p_1, \dots p_n)$ be a function defined for any integer n and for all values $p_1, p_2, \dots p_n$ such that $p_i \geq 0 \ \forall\, i$ and $\sum_i p_i = 1$. If for any function, we have this function to be continuous w.r.t. all its arguments and if it satisfies prop. 1, 2, & 3, then

$$H(p_1, p_2, \dots p_n) = -\lambda \sum_{k=1}^{n} p_k \log p_k$$

(λ is +ve constant)

<u>Proof</u>:

① To show $L(n) = f\left(\frac{1}{n}, \frac{1}{n}, \dots \frac{1}{n}\right) = +\lambda \log n$.

from i) & ii) we have

$$L(n) = f\left(\frac{1}{n}, \frac{1}{n}, \dots \frac{1}{n}\right) = f\left(\frac{1}{n}, \frac{1}{n}, \dots \frac{1}{n}, 0\right) \leq f\left(\frac{1}{n+1}, \frac{1}{n+1}, \dots \frac{1}{n+1}\right) = L(n+1)$$

$\Rightarrow$ $L(n) \leq L(n+1) \rightarrow L(n)$ is non-decreasing function of n

Let m and r be 2 integers,

Consider m mutually ind. prob. dist. $S_1, S_2, \dots S_m$ each of which contain r equally likely events.

Tut

① X = time to leave building

$E[X] = ?$

$A \to 8$, $B \to 4$, $C \to 3$

$P(X=1) = 0$

$2 \Rightarrow 0$

$3 = \left(\frac{1}{3}\right)$

$4 = 0$

$5 : 0$

$6 : 0$

$x^2 + y^2 \le k$

$\sum_x \frac{1}{\alpha^x} \sum_y \frac{1}{\alpha^y}$

$\frac{1}{\alpha - 1} = 1$

$\alpha - 1 = 1$

$\boxed{\alpha = 2}$

$\frac{1}{2}$

$\frac{1}{4} \quad \frac{1}{4}$

$\frac{1}{9} \quad \frac{1}{9}$

$\sum_x e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \lambda$

$\text{mean} = \sum_{x=0}^{\infty} x \cdot e^{-\lambda} \frac{\lambda^x}{x!} =$

→ $f(S_i) = f\left(\frac{1}{r}, \frac{1}{r}, \cdots \frac{1}{r}\right) = L(r). \quad [1 \le i \le m]$

Now $f(S_1 S_2 S_3 \ldots S_m) = \sum_{i=1}^{m} f(S_i) = m \cdot L(r).$ — (I)

Now we can see that $\{S_1, S_2, \ldots S_m\}$ contains a total of r^m equally likely events.

Thus, $f(S_1, S_2, \ldots S_m) = L(r^m)$ — (II)

using eq (I) & (II). $\boxed{L(r^m) = m \cdot L(r)}$ — (III)

Similarly for n and s $\boxed{L(s^n) = n \cdot L(s)}$ — (IV) let r, s, n be arbitrary and m be such that

$$r^m \leq s^n \leq r^{m+1}$$

$$m \log r \leq n \log s \leq (m+1) \log r \qquad \text{(since log is inc. function)}$$

$$\Rightarrow \boxed{\frac{m}{n} \leq \frac{\log s}{\log r} \leq \frac{m+1}{n}} \text{ ——— } A$$

also, since L is an inc. function (proved earlier),

$$L(r^m) \leq L(s^n) \leq L(r^{m+1}).$$

$$m \cdot L(r) \leq n \cdot L(s) \leq (m+1) L(r).$$

$$\boxed{\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m+1}{n}} \text{ ——— } B$$

$$\frac{m}{n} \leq \frac{\log s}{\log r} \leq \frac{m}{n} + \frac{1}{n} \quad \text{and} \quad \frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m}{n} + \frac{1}{n}. \text{ — eq 2}$$

$$-\frac{m}{n} - \frac{1}{n} \leq \frac{-\log s}{\log r} \leq -\frac{m}{n} \text{ — eq 1}$$

adding eq 1 & eq 2 $\quad -\frac{1}{n} \leq \left(\frac{L(s)}{L(r)} - \frac{\log s}{\log r} \right) \leq \frac{1}{n}$

$$\Rightarrow \boxed{\left| \frac{L(s)}{L(r)} - \frac{\log s}{\log r} \right| \leq \frac{1}{n}}$$

Now since we chose arbitrary n, we can say $n \to \infty$

$$\Rightarrow \frac{L(s)}{L(r)} = \frac{\log s}{\log r} = \lambda \text{ (say)} \Rightarrow L(s) = \lambda$$

$$\frac{L(s)}{\log s} = \frac{L(r)}{\log r} = \lambda \text{ (say)} \Rightarrow L(s) = \lambda \log(s).$$

In general $\quad L(n) = \lambda \log n$

Set $p_i = \frac{g_i}{g}$ such that $\sum_i g_i = g$. $g_i \to$ positive integer

Let A be random variable with dist $A_i = p_i \quad \forall\, i \in [1, n]$

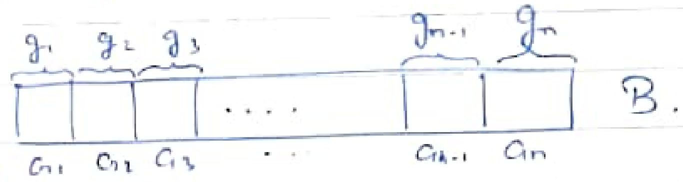
Let B be a random variable with dist. $B = \{B_1, B_2, \ldots B_g\}$ events

Here B is dependent on A.

we need to check uncertainity measure for A.

$$f(A) = -\lambda \sum_{i=1}^{n} p_i \log p_i \longrightarrow \text{check.}$$

Let us group B as



If the event A_i occured in A, then in the scheme B, all the g_i events in group G_i will have prob. $\frac{1}{g_i}$ and rest of the events will have prob. 0.

$$f_i(B) = f\left(\frac{1}{g_i}, \frac{1}{g_i}, \cdots \frac{1}{g_i}\right) = L(g_i) = \lambda \log(g_i).$$

Now $f_A(B) = \sum_{i=1}^{n} p_i f_i(B) = \lambda \sum_{i=1}^{n} p_i \log g_i \}$ ~~average uncertainity of scheme A.~~

The uncertainity of scheme B given A has occured.

$$f_A(B) = \lambda \sum_{i=1}^{n} p_i \log(g_i). \quad \text{—— (1).}$$

$A_k B_l$ be joint dist. $[1 \le i \le n,\ 1 \le l \le g]$

$$\Rightarrow P(A_k B_l) = P(A_k) \cdot P(B_l \mid A_k) = p_i \cdot \left(\frac{1}{g_i}\right) = \frac{1}{g}$$

Thus AB is uniform dist.

$$f(AB) = f(A) + f_A(B)$$

$$\lambda \log g = f(A) + \lambda \sum_{i=1}^{n} p_i \log(p_i g).$$

$$= f(A) + \lambda \sum_{i=1}^{n} p_i \log p_i + \lambda \log g \left(\sum_{i=1}^{n} p_i\right)$$

$$f(A) = -\lambda \sum_i p_i \log p_i$$

$S \to$ set.

$T \subseteq S$ $\quad$ $C_S(T) \to$ relative compliment of T in S

$C_S(T) = \{x : x \in S \text{ and } x \notin T\}$

Let S be a system (collection) of sets.

& $U \in S$

if $\forall A \in S$, $\underbrace{A \cap U = A}_{\text{unit}}$

Let X be a set and Σ be system/collection of subsets of X

$\Sigma \leftarrow \sigma$ algebra iff Σ satisfies

1.) $X \in \Sigma$ (unit)

ii) $\forall A \in \Sigma$, $C_X(A) \in \Sigma$

iii). $\forall A_n \in \Sigma$, $\bigcup_{n=1}^{1,2,\ldots\infty} A_n \in \Sigma$

$(X, \Sigma) \to$ measurable space

$(X, \Sigma, \mu) \to$ Measure space prob

$(\Omega, \Sigma, P_x) \to$ Measure Space

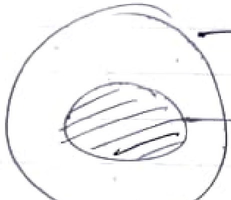
Ω → sample space, Σ → event space, P_x → map.

MISSED 1 CLASS.

Consequences of Asymptotic Equipartition Theorem.

Typical Set : $T_\epsilon^{(n)}$

Atypical set : $[T_\epsilon^{(n)}]^c$



$\rightarrow \mathcal{X}^n = |\mathcal{X}|^n$ (set of all sequences of length n)

$\rightarrow$ Typical set $\quad T_\epsilon^{(n)} = |T_\epsilon^{(n)}|$

we have shown, $\quad (1-\delta)\, 2^{n(H(X)-\epsilon)} \leq |T_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$

since $\quad |T_\epsilon^{(n)}|$ is bounded by $2^{n(H(X)+\epsilon)}$.

they require atmost $n(H(X)+\epsilon)$ bits to encode $T_\epsilon^{(n)}$

* Let us assume that $X_1, X_2, \dots X_n$ are i.i.d. r.v. with pmf $p(x)$.
We divide all sequences in $\mathcal{X}^n$ in two sets.(Typical set $\rightarrow T_\epsilon^{(n)}$)
and (ATypical set $\rightarrow [T_\epsilon^{(n)}]^c$)

We know that $|T_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

If we maintain a particular order of indexing. it requires no more than $\boxed{n(H(X)+\epsilon)+1 \text{ bits.}}$

For ATypical set, you do not require more than
$\boxed{n \log |\mathcal{X}| + 1}$ bits

to be able to distinguish b/w Typ. & ATyp,

0 → Typical → $n(H(x)+\epsilon)+2$

1 → Atypical → $n \log |\mathcal{X}| + 2$.

Let $l(x^n)$ be the length of codeword corresponding to sequence x^n

$$\mathbb{E}[l(x^n)] = \sum_{x^n \in \mathcal{X}^n} p(x^n) \cdot l(x^n)$$

$$= \sum_{x^n \in T_\epsilon^n} p(x^n)\, l(x^n) + \sum_{x^n \in (T_\epsilon^n)^c} p(x^n) \cdot l(x^n).$$

$$\leq \sum_{x^n \in T_\epsilon^n} p(x^n) \cdot [n(H(X)+\epsilon)+2] + \sum_{x^n \in [T_\epsilon^n]^c} p(x^n)[n \log|\mathcal{X}| + 2]$$

$$= [n(H(X)+\epsilon)+2] \underbrace{\sum_{x^n \in T_\epsilon^n} p(x^n)}_{P\{T_\epsilon^n\}} + [n \log|\mathcal{X}|+2] \underbrace{\sum_{x^n \in [T_\epsilon^n]^c} p(x^n)}_{P\{[T_\epsilon^n]^c\}}$$

$$= \Pr\{T_\epsilon^n\}[n(H(X)+\epsilon)] + \Pr\{[T_\epsilon^n]^c\} \cdot [n \log|\mathcal{X}|] + 2\{\underbrace{\Pr\{T_\epsilon^n\} + \Pr\{[T_\epsilon^n]^c\}}_{1}\}$$

$$= \Pr[T_\epsilon^n][n(H(X)+\epsilon)] + \Pr[[T_\epsilon^n]^c][n \log|\mathcal{X}|] + 2$$

Now we know $\Pr[T_\epsilon^n] \geq 1-\delta \Rightarrow \Pr[(T_\epsilon^n)^c] \leq \delta$

also $\Pr[T_\epsilon^n] \leq 1$

Thus,

$$\mathbb{E}[l(x^n)] \leq n(H+\epsilon) + \delta n \log|\mathcal{X}| + 2$$

$$\leq nH + n(\underbrace{\epsilon + \delta \log|\mathcal{X}| + 2/n}_{\epsilon'})$$

$$\leq nH + n\epsilon'$$

$$\mathbb{E}[l(x^n)] \leq n(H+\epsilon')$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{n} l(x^n)\right] \leq H + \epsilon'$$

" The average length of the sequence is bounded by Shannons entropy."

OR x^n can be represented by using $nH(X)$ bits on average

Q) If n is not very large, which set will be bigger, Typ or ATyp.?

▶ Markov Process

Eg Let P_{ij} = prob. of transitioning from state i to state j
$i \in \{0,1\}$, $j \in \{0,1\}$.

if it rains today (0) → rains tom (α)
If it does not rain today (1). → " " (β)

$P_{00} = \alpha$ $P_{01} = 1-\alpha$
$P_{10} = \beta$ $P_{11} = 1-\beta$

current state \ next	0	1
0	α	$1-\alpha$
1	β	$1-\beta$

* Prob of something happening in the future is only dep. on present and not the past.
* Pdf of state i is only dep on state i-1

A stochastic process $\{X_n : n = 0, 1, 2, \ldots\}$ takes on a finite number (countable) of values

Let $X_1, X_2, \ldots X_n$ be the r.v.

$$\Pr\{(X_1, X_2, \ldots X_n) = (x_1, x_2, \ldots x_n)\} = p(x_1, x_2, \ldots x_n)$$
$$(x_1, x_2, \ldots x_n) \in \mathcal{X}^n \text{ for } n = 1, 2, \ldots$$

A stochastic process is called stationary if the joint distribution of any ~~sequence of~~ subset of the sequence of random variable is invariant w.r.t. time shifts.

$$\Pr\{X_1 = x_1, X_2 = x_2 \ldots X_n = x_n\}$$
$$= \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots X_{n+l} = x_n\}$$
(for any shift l).

A discrete ~~stho~~ stochastic process will be called as Markov process for $n = 1, 2, \ldots$ if

$$\Pr(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1} \ldots X_0 = x_0)$$
$$= \Pr(X_{n+1} = x_{n+1} \mid X_n = x_n)$$
$$\forall (x_1, x_2, \ldots x_n) \in \mathcal{X}^n$$

Entropy Rule for Stochastic Processes

A stochastic process is a sequence of indexed r.v.

→ The joint pmf.

$$\Pr((X_1, X_2, \ldots X_n) = (x_1, \ldots x_n))$$
$$= p(x_1, x_2, \ldots x_n)$$
$$(x_1, x_2 \ldots x_n) \in \mathcal{X}^n$$
$$\{0,1\} \qquad \{0,1\}^n$$

Stationary: If joint distribution of any subset of the sequence is invariant w.r.t. time (or time shifts).

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = P(X_{\ell+1} = x_1, X_{\ell+2} = x_2, \ldots X_{\ell+n} = x_n).$$

for any shift ℓ

A discrete ~~sche~~ stochastic process will be known as Markov Process if for $n = 1, 2, \ldots$

$$P(\underbrace{X_{n+1} = x_{n+1}}_{\text{future}} \mid \underbrace{X_n = x_n}_{\text{present}}, \underbrace{X_{n-1} = x_{n-1} \ldots X_1 = x_1}_{\text{past}}) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

OR $$P(x_1, x_2, \ldots x_n) = p(x_1) \cdot p(x_2 | x_1) \cdot p(x_3 | x_2) \ldots p(x_n | x_{n-1}).$$

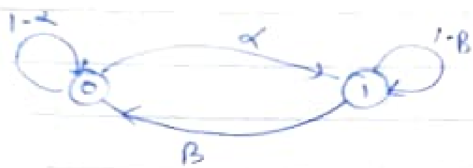
Stationary Markov Process:

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n) = P(X_2 = x_2 \mid X_1 = x_1).$$

Transition Probabilities : If $\{X_i\}$ is a Markov chain, then X_n represents the state at time n

$$P_{ij} = \Pr\{X_{n+1} = j \mid X_n = i\} \qquad \{i, j = 1, 2, \ldots n$$

if MP is stationary then P_{ij} remains same for all X_k



$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

Let $[\mu_1\ \mu_2]$ be the prob of being in S_0 and S_1 respectively, then $\mu = [\mu_1\ \mu_2]$ is invariant (due to stationary MP).

$$[\mu_1\ \ \mu_2] \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = [\mu_1\ \ \mu_2]$$

$$\Rightarrow \left.\begin{aligned} \mu_1(1-\alpha) + \mu_2\beta &= \mu_1 \\ \mu_1\alpha + \mu_2(1-\beta) &= \mu_2 \end{aligned}\right\} \Rightarrow \boxed{\mu_1\alpha = \mu_2\beta} \text{ — (1)}$$

Note that $\underbrace{\mu_1 + \mu_2 = 1}_{(2)} \Rightarrow \boxed{\mu_1 = \dfrac{\beta}{\alpha+\beta} \qquad \mu_2 = \dfrac{\alpha}{\alpha+\beta}}$

$$H(X_n) = -\left(P(X_n=1).\log P(X_n=1) + P(X_n=0).\log P(X_n=0)\right)$$

$$= \frac{\alpha}{\alpha+\beta}\log\frac{\alpha+\beta}{\alpha} + \frac{\beta}{\alpha+\beta}\log\frac{\alpha+\beta}{\beta}$$

$$= \log(\alpha+\beta).1 - \frac{(\alpha\log\alpha + \beta\log\beta)}{\alpha+\beta}.$$

* If we can go from any state of a Markov chain to any other ~~chain~~ state with positive probability in finite steps, the the Markov chain is called irreducible.

▶ Chapman-Kolmogorov Eqⁿ

One step transition Prob $\longrightarrow P_{ij}$

$P_{ij}^n \rightarrow$ Prob of reaching j from i in n steps.

- $P_{ij}^n = \Pr\{X_{n+m}=j \mid X_m = i\} \quad n \geq 0, \quad i,j \geq 0$

$$P_{ij}^{n+m} = \Pr\{X_{n+m}=j \mid X_0=i\}. = \sum_{k=0}^{\infty} \Pr\{X_{n+m}=j, X_n = k \mid X_0 = i\}.$$

$$= \sum_{k=0}^{\infty} P\{X_{n+m}=j \mid X_n = k, X_0 = i\}.\ P\{X_n = k \mid X_0 = i\}.$$

$$= \sum_{k=0}^{\infty} P_{kj}^m \cdot P_{ik}^n$$

Entropy Rate of Stochastic Process

$\lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \dots X_n) \leftarrow H(\mathcal{X})$

↳ Stochastic process.

another quantity $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1} \dots X_1)$

↳ if lim exists

Thm: For any stochastic process, this limit exists.

& $H(\mathcal{X}) = H'(\mathcal{X})$

Lemma 1: If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$ then $b_n \to a$ for large n

Proof 1

$\Rightarrow$ ~~$b_n = a$~~ $\lim_{n\to\infty} b_n = \lim_{n\to\infty} \frac{1}{n} \sum_1 a_i = \lim \frac{\sum_{i=1}^{n} \left(\lim_{n\to\infty} a_i\right)}{n} = \lim \frac{\cancel{n} a}{\cancel{n}} = a$

Proof 2

we know $|a_n - a| < \epsilon$ for $n > N(\epsilon)$.

Now $|b_n - a| = \left| \frac{1}{n} \sum_1 a_i - a \right| = \left| \frac{1}{n} \sum (a_i - a) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |a_i - a|$

$= \frac{1}{n} \sum_{i=1}^{N(\epsilon)-1} |a_i - a| + \frac{1}{n} \sum_{i=N(\epsilon)}^{n} |a_i - a|$

$= \frac{1}{n} \sum_{i=1}^{N(\epsilon)-1} |a_i - a| + \frac{(n - N(\epsilon))}{n} \cdot \epsilon$

$\frac{n - N(\epsilon)}{n} < 1$

$\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)-1} |a_i - a| + \epsilon \;\forall\; n \geq N(\epsilon)$

$< \epsilon$ for large n

$\Rightarrow |b_n - a| < 2\epsilon$ for large n.

Hence $b_n \to a$ as $n \to \infty$

Recap

$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \dots X_n)$ → Entropy Rate for a stochastic process

↳ when the limit exists.

$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1} \dots X_1)$.

(when the limit exists).

Thm: If we have a stationary stochastic process $\mathcal{X}$, the limit exists and H and H' are equal $\Rightarrow H(\mathcal{X}) = H'(\mathcal{X})$.

Lemma 2: For a stationary stochastic process $H(X_n | X_{n-1} \dots X_1)$ is decreasing in n and has a limit $H'(\mathcal{X})$

We know. $H(X_{n+1} | X_n X_{n-1}, \dots X_1) \le H(X_{n+1} | X_n \dots X_2)$

$= H(X_n | X_{n-1} \dots X_1)$ (Because stationary)

$\Rightarrow H(X_{n+1} | X_n \dots X_1) \le H(X_n | X_{n-1} \dots X_1)$

So this quantity is decreasing as well as non negative

↳ $H'(\mathcal{X})$ exists (limit exists)

Now using Lemma1 & Lemma 2, we prove Thm.

→ We have $\frac{1}{n} H(X_1, \dots X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1} \dots X_1)$. [chain Rule]

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n).$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \underbrace{H(X_i | X_{i-1} \dots X_1)}_{\text{tends to } H'(\mathcal{X})}.$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1} \dots X_1) \to H'(\mathcal{X})$$

[using Lemma 1].

$\Rightarrow \boxed{H(\mathcal{X}) = H'(\mathcal{X})}$

~~Find~~

* Find entropy rate of SMC

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2 \dots X_n) = H'(X)$$

$$= \lim_{n\to\infty} H(X_n | X_{n-1} \dots X_1) = \lim_{n\to\infty} H(X_n | X_{n-1})$$

$$= \lim_{n\to\infty} H(X_2 | X_1) = H(X_2|X_1)$$

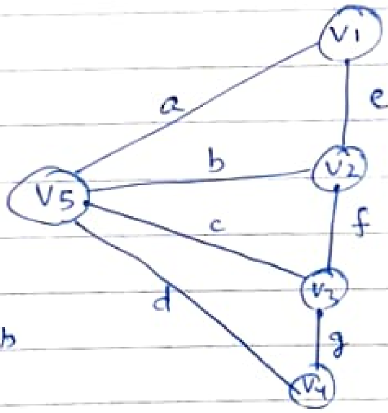
$\Rightarrow$ Entropy Rate of SMC is $H(X_2|X_1)$

If $P_{ij} = P$ and stationary dist $\mu = \{\mu_i\}$ $\mathcal{X} = \{X_i\}$ (SMC).

$$\Rightarrow H(\mathcal{X}) = H(X_2|X_1) = -\sum_{ij} \mu_i P_{ij} \log(P_{ij})$$

Entropy Rate of a Random Walk on a Weighted Graph.

$\{V_1, V_2, \dots V_5\}$
$\{a, b, c, \dots g\}$.



$W_{V_5} = a+b+c+d$
$W_{V_4} = d+g$
$W_{V_3} = g+f+c$
$W_{V_2} = e+b+f$
$W_{V_1} = a+e$

$\sum W_{V_i} = 2W$
W = sum of edge weights

Now consider a graph having m nodes $\{1, 2, 3, \dots m\}$. with weights $W_{ij} \geq 0$ (connecting node i to node j) & Assuming $W_{ij} = W_{ji}$ (undirected graph).
* we call $W_{ij} = 0$ if i, j are not connected.

Consider a random walk on graph. $\{X_n\}$, $X_n \in \{1, 2, \dots m\}$.

- $$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}$$

Now let $\{X_n\}$ be SMC.

Let $w_i = \sum_j w_{ij}$ = total weights of edges coming out of i

Let $w = \sum_{ij} w_{ij}$ = total weight of all the edges

$$\Rightarrow \sum_i w_i = 2w$$

Now we know. $\mu_j = \sum_i \mu_i . P_{ij}$ —— (1)

let $\mu_i = \frac{w_i}{2w} \Rightarrow$ In (1),

$$\sum_i \mu_i P_{ij} = \sum_i \frac{w_i}{2w} \cdot \frac{w_{ij}}{\sum_k w_{ik}} = \sum_i \frac{\cancel{w_i}}{2w} \frac{w_{ij}}{\cancel{w_i}} = \frac{w_i}{2w} \cdot \mu_i$$

Now $H(\mathcal{X}) = H(X_2 | X_1) =$

$$= -\sum_i p(X_2) . \sum_j p(x_2|x_1) \log(p(x_2|x_1)).$$

$$= -\sum_i \mu_i \sum_j P_{ij} \log(P_{ij})$$

$$= -\sum_i \frac{w_i}{2w} \cdot \sum_j \left(\frac{w_{ij}}{w_i}\right) \cdot \log\left(\frac{w_{ij}}{w_i}\right)$$

$$= -\sum_i \sum_j \frac{\cancel{w_i}}{2w} \frac{w_{ij}}{\cancel{w_i}} \log\left(\frac{w_{ij}}{w_i}\right) = -\sum_i \sum_j \frac{w_{ij}}{2w} \log\left(\frac{w_{ij}}{w_i}\right)$$

in prev eg. $H(\mathcal{X}) = \frac{-1}{2w}\Big[a \log\left(\frac{a}{a+b+c+d}\right) + a \log\left(\frac{a}{a+e}\right)$

$+ b \log\left(\frac{b}{a+b+c+d}\right) + b \log\left(\frac{b}{b+e+f}\right)$

$+ c \log\left(\frac{c}{a+b+c+d}\right) + c . \log\left(\frac{c}{f+g+c}\right)$

$+ d \log\left(\frac{d}{a+b+c+d}\right) + d \log\left(\frac{d}{g+d}\right)$

$+ e \log\left(\frac{e}{e+a}\right) + e \log\left(\frac{e}{e+b+f}\right)$

$+ f \log\left(\frac{f}{c+f+g}\right) + f \log\left(\frac{f}{b+e+f}\right)$

$+ g \log\left(\frac{g}{d+g}\right) + g \log\left(\frac{g}{c+f+g}\right)\Big]$

$\frac{-1}{2w}\Big[2[a\log a + b\log b + c\log c + d\log d + e\log e + f\log f + g\log g]$

$- \log(a+b+c+d) - \log(a+e) - \log(b+e+f) - \log(g+d) - \log(g+c+f)$

▷ Hidden Markov Models

$\{X_1, X_2 \dots X_n\} \rightarrow$ SMC

$\{Y_1, Y_2 \dots Y_n\} \rightarrow$ such that $y_i = \phi(x_i)$

~~$H(y) = \lim_{n\to\infty} \frac{1}{n} H(Y_1, Y_2, \dots Y_n)$~~

~~$= \lim_{n\to\infty} \frac{1}{n} H($~~ $\qquad Y_{n+k}$

Lemma:

~~$H(X_n | X_{n-1} \dots X$~~

$H(Y_n | Y_{n-1} \dots Y_2 X_1) \leq H(Y)$.

$\hookrightarrow = H(Y_n | Y_{n-1} \dots Y_2\, Y_1\, X_1)$ since $\phi(X_1) = Y_1$

$= H(Y_n | Y_{n-1} \dots Y_2, Y_1\, X_1, X_0, X_{-1} \dots X_{-k})$. ($X$ is SMC)

$= H(Y_n | Y_{n-1} \dots Y_1\, X_1\, X_0 \dots X_{-k}, Y_0 \dots Y_{-k})$ $\quad y_i = \phi(x_i)$

$\leq H(Y_n | Y_{n-1} \dots Y_1\, Y_0 \dots Y_{-k})$ (Conditioning Reduces Entropy)

$= H(Y_{n+k-1} | Y_{n+k} \dots Y_1)$

$\hookrightarrow$ (Y is stationary, $Y_i = \phi(X_i)$)

This is true $\forall$ k.

putting limit k

$\lim_k H(Y_{n+k-1} | Y_{n+k} \dots Y_1) = H(Y)$

$H(Y_n | Y_{n-1} \dots Y_2\, Y_1) \leq \lim_k H(Y_{n+k-1} | Y_{n+k} \dots Y_1)$

$\hookrightarrow H'(y) = H(y)$

$$\Rightarrow \boxed{H(Y_n | Y_{n-1}\, Y_{n-2} \dots Y_2\, X_1) \leq H(Y) \leq H(Y_n | Y_{n-1} \dots Y_1)}$$

Proof:

$$\Delta = -H(Y_n \mid Y_{n-1} \cdots X_1) + H(Y_n \mid Y_{n-1} \cdots Y_1 X_1).$$

$$\Delta = I(X_1 ; Y_n \mid Y_{n-1} \cdots Y_1).$$

Now is $I(X_1 ; Y_1 Y_2 \cdots Y_n) \leq H(X)$?

$\hookrightarrow$

$$\lim_{n\to\infty} I(X_1 ; Y_1 Y_2 \cdots Y_n) \leq H(X_1)$$

by chain rule

~~$\lim_{n\to\infty} I(X_1, Y_2, \ldots$~~

$$\lim_{n\to\infty} I(X_1 ; Y_1, Y_2, \ldots Y_n) = \lim_{n\to\infty} I(X_1 ; Y_i \mid Y_{i-1} \cdots Y_1)$$

$$= \sum_{i=1}^{\infty} I(X_1 ; Y_i \mid Y_{i-1} \cdots Y_1)$$

infinite sum → finite — terms are non negative

$\Rightarrow$

the sequence $\{ I(X_1 : Y_n \mid Y_{n-1} \cdots Y_1) \}$ converges to 0.

OR $\lim_{n\to\infty} I(X_1 ; Y_n \mid Y_{n-1} Y_{n-2} \cdots Y_1) \to 0$

$$\Rightarrow \boxed{\lim_{n\to\infty} \Delta = 0}$$

A source code C corresponding to a r.v. X is a mapping
$C: \mathcal{X} \longrightarrow D^* \subset \{0,1\}^*$.

{Range of R.V. X}   {set of finite strings of symbols taken from D-any alphabet}.

The expected length $L(C)$ of a particular code C corresponding to a r.v. X which is occurring with pmf $P(x)$

$$L(C) = \sum_{x \in \mathcal{X}} p(x) \cdot l(x)$$

$$L(C) = \frac{4}{12} \cdot 1 + \frac{2}{12} \cdot 2 + \frac{2}{12} \cdot 2 + \frac{2}{12} \cdot 2 + \frac{1}{12} \cdot 3 + \frac{1}{12} \cdot 3$$

$$= \frac{4+4+4+4+6}{12} = \frac{22}{12} = \frac{11}{6}$$

$$H(X) = \frac{4}{12} \cdot \log \frac{12}{4} + \frac{2 \times 3}{12} \cdot \log \frac{12}{2} + \frac{2 \times 1}{12} \cdot \log 12$$

$$= \frac{1}{3} \log 3 + \frac{1}{2} \log 6 + \frac{1}{6} \log 12$$

$$\frac{1}{3} \cdot 1.58 + \frac{1}{2} \cdot 2.58 + \frac{1}{6} \cdot 3.58$$

$$0.526 + 1.29 + 0.596$$

$$\longrightarrow 2.41.$$

		$l(x)$	$P(x)$
a	0	1	4/12
b	01	2	2/12
c	10	2	2/12
d	11	3	2/12
e	110	3	2/12
f	101	3	1/12
			1/12

$$L(C) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 \times 2$$

$$\frac{4+4+6}{8} = \frac{14}{8} = \frac{7}{4}$$

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + 2 \times \frac{1}{8} \cdot 3 = \frac{4+4+6}{8} = \frac{7}{4}$$

Non - Singular Code

A code will be called as non-singular code if different elements of the range $\mathcal{X}$ of the r.v X is mapped into different strings in D^*.

$$x_i \neq x_j \Rightarrow c(x_i) \neq c(x_j).$$

The extension of C is C^* is defined as the mapping of finite length strings of $\mathcal{X}$ to finite length strings of D.

x_1	0
x_2	0
x_3	0
x_4	1

singular

x_1	0
x_2	1
x_3	011
x_4	01

Non singular

$$C(x_1 . x_2, \dots x_n) = c(x_1).c(x_2)\cdots c(x_n).$$

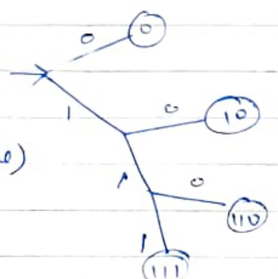
$\hookrightarrow$ concatenation of codewords.

A code is called uniquely decodable if its extension is non-singular.

x	C	C'
x_1	0	10
x_2	010	00
x_3	01	11
x_4	10	1101

C is not uniquely decodable since $c(x_1 x_4) = c(x_3 x_1) = c(x_2) = 010$

Krafts Inequality



Given a prefix code (instaneous code) over an alphabet D (size)., the codewords of different lengths $l_1, l_2, \dots l_m$ must satisfy the inequality

$$\sum_i D^{-l_i} \leq 1$$

Given a set of codewords lengths satisfying this Inequality then there exists an instaneous code ~~with codeword length~~.

Let l_{max} be the length of the longest codeword.

Consider all the nodes in the tree at the level l_{max}.

→ some of them are codewords,

→ some of them are descendants

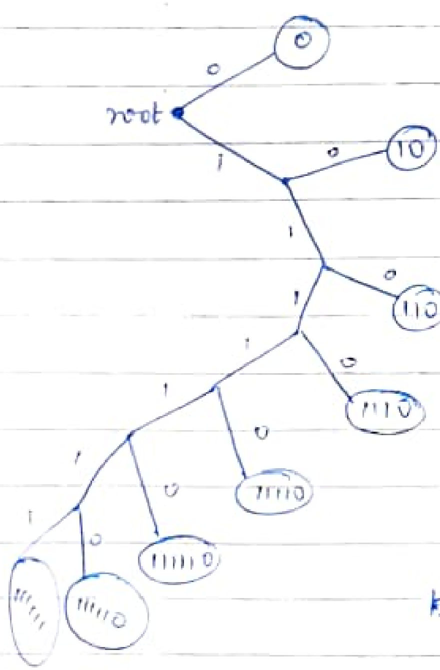
A codeword at level l_i has $D^{l_{max} - l_i}$ descendents at the level l_{max}

Now also, no. of descendants in l_{max} ≤ total possible no of strings in l_{max}

$$\sum_i D^{l_{max} - l_i} \le D^{l_{max}}$$

$$\Rightarrow \boxed{\sum_i D^{-l_i} \le 1}$$

X	P(x)	c(x)	l(x)
1	1/2	0	1
2	1/4	10	2
3	1/16	110	3
4	1/16	1110	4
5	1/32	11110	5
6	1/32	111110	6
7	1/32	1111110	7
8	1/32	1111111	7



$\boxed{D=2}$

$$\sum D^{-l_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-7} + 2^{-7}$$

$$= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} \; \frac{1}{32} \; \frac{1}{16}$$

$$= 1 \le 1$$

~~H(x)~~ $L(c) = \sum l(x).p(x)$

$$= \frac{1}{2} + \frac{1}{2} = \frac{16 + 16 + 6 + 8 + 5 + 6 + 14}{32}$$

$$L(c) = \frac{71}{32}$$

$\frac{17}{8}$ $\frac{68}{32}$ $H(X) = \frac{1}{2} + \frac{2}{4} + \frac{4}{16} \times 2 + \frac{4 \times 5}{32}$

$$\boxed{H(x) \le L(C) < H(x) + 1}$$

Optimal Code

Any code satisfying prefix code must satisfy Kraft's inequality.
↳ sufficient condition for the existence of a code (unordered set) from a given set of lengths of codeword.

Target: Find the prefix code with min. expected length.

$C \to$ prefix code / prefix free
$L(C) \to$ Expected/average length (minimize). $L(C) = \sum p(x_i).l(i)$
s.t. $\sum D^{-l_i} \le 1$

By method of Langrange's multiple

$$J = \sum p(x_i).l_i + \lambda\left(\sum D^{-l_i} - 1\right)$$

$$\frac{\partial J}{\partial l_k} = p(x_k) - \lambda D^{-l_k} \ln D = 0$$

$$\Rightarrow D^{-l_k} = \frac{p(x_k)}{\lambda \ln D} \quad \text{--- (1)}$$

Now $\sum_k D^{-l_k} \le 1 \Rightarrow \sum_k \frac{p(x_k)}{\lambda \ln D} \le 1 \Rightarrow \left.\frac{1}{\lambda \ln D} \le 1 \right\} \text{--- (2)}$

$\therefore D^{-l_k} = \frac{p(x_k)}{\lambda \ln D} \le p(x_k)$. (from (2)).

$\Rightarrow D^{-l_k} \le p(x_k)$.

$$\boxed{l_k \ge -\log_D p(x_k).}$$

The average exp. length of instantaneous code

$$L = \sum_i p(x_i) l_i \ge -\sum p(x_i) \log_D p(x_i)$$

$$L(C) \ge H_D(x) = \frac{H(x)}{\log D}$$