Introduction to Information Theory.

Info theory → Algorithmic
Info theory → Theory.

→ What is information?

Information → A numerical measure of the uncertainty of an experimental outcome.

More the random, more the information.

→ How to quantify information?
→ How can we measure the amount of info?
→ ensure correctness of info?

Information Theory

→ Information theory was invented by Shannon in 1948.
→ One of the basic postulates of information theory is that information can be treated like a measurable physical quantity, such as density, or mass.
→ The information content of a message consists simply of the numbers 0's and 1's it takes to transmit it. Information can thus be measured in bits.

Basic concepts in info theory.

- Entropy
- Joint Entropy
- Conditional Entropy

Shannon Info content

→ we study info content of those events to which we can associate probability of occurence of these events.

Eg: $P(A) = 0.5$

$P(B) = 0.125$

→ $P(B)$ provides more information

Shannon information content $h(A)$ corresponding to event A is defined as.

$$h(A) = \log_2\left(\frac{1}{P(A)}\right) \text{ bits.}$$

← info associated with $P(A)$

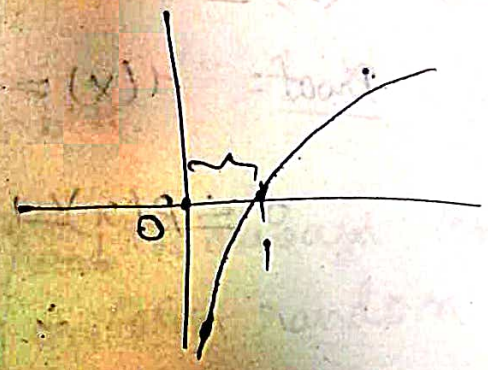
Information

$$h(A) = \log_2\left(\frac{1}{(1/2)}\right)$$

$$= \log_2 2 = 1$$

$$h(B) = \log_2\left(\frac{1}{1/8}\right)$$

$$= \log_2 8 = 3$$



we extend this definition to probability distribution function.

X → discrete r.v.

Eg $X = \{1, 2, 3\}$

$P(1)$ $P(2)$ $P(3)$

Entropy is like the average information content.

$$H(X) = \sum_{x \in X} p(x) \log\left(\frac{1}{p(x)}\right)$$

Entropy.

→ If base to logarithm is 2, then entropy is expressed in bits → default

→ If base is to log is e, then entropy is expressed in nats.

→ If base to log is b, then we denote entropy by $H_b(X)$.

Interpretation of $H(X)$ in terms of expectation:

$$H(X) = E_p \log\left(\frac{1}{p(x)}\right) \quad \text{where } X \sim p(x)$$

Eg. Properties of $H(X)$

1.) $H(X) \geq 0$

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

$0 \leq p(x) \leq 1$ so $\log(p(x)) \leq 0$

$\therefore H(X) \geq 0$

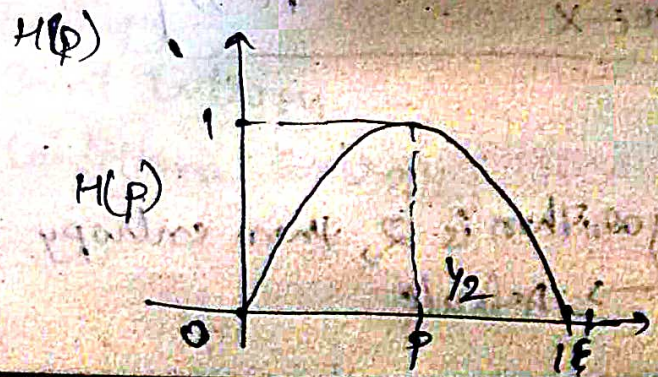
2.) $H_b(X) = (\log_b a)\, H_a(X)$

Eg. Entropy of a Bernoulli r.v

Bernoulli r.v with parameter p:

Entropy of Bernoulli r.v

$$= -p \log_2 p - (1-p) \log_2 (1-p) =: H(p)$$



$H(p)$ is always a concave function.

Entropy of a fair coin = 1 bit.

Entropy of a discrete uniform r.v

Eg. consider a uniform r.v with support set $\{1, 2, \ldots, n\}$.

$$H(X) = -\frac{1}{n} \times \log\left(\frac{1}{n}\right) - \frac{1}{n} \times \log\left(\frac{1}{n}\right) - \ldots \text{ } n \text{ times}$$

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

$\therefore H(X) = \log n$

Eg. consider a fair dice.

$H(X) = \log_2 6 = 2.585$ bits.

Entropy of any dice (biased/~~unbiased~~) is always less than or equal to entropy of uniform random variable.

1	2	3	4	5	6
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0

$$H(X) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 3/2 = 1.5$$

$\therefore$ $\boxed{H(X) \leq \log |X|}$

Entropy of uniform r.v

where $|X|$ denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over X.

Joint entropy & conditional Entropy

→ Joint entropy $H(X,Y)$ of a pair of discrete r.vs (X,Y) is defined as.

$$\boxed{H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log(p(x,y))}$$

$$\boxed{H(X,Y) = -E\log p(X,Y)}$$

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

$$H(Y|X) = \sum_{x\in X} p(x) H(Y|X=x)$$

$$= \sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log(p(y|x))$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

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

E → expectation with respect to joint r.v and not conditional r.v

$$E(F(X)) = \sum_x F(x)\,p(x)$$

$$E(F(X,Y)) = \sum_x \sum_y F(x,y)\,p(x,y)$$

Entropy ≡ uncertainty

What is $H(X,Y)$ if X & Y are independent?

X Y

H(X) H(Y)

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

$H(X/Y)$ is not necessarily equal to $H(Y/X)$.

Relation between $H(X,Y)$ and $H(X/Y)$ is

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

↗ called the chain rule

$$\boxed{\begin{aligned} H(X,Y) &= H(X) + H(Y/X) \\ &= H(Y) + H(X/Y) \end{aligned}}$$

Proof:

$$\text{LHS} = H(X,Y)$$
$$= -\sum_x \sum_y p(x,y) \log\left(p(x,y)\right)$$

$$p(x,y) = p(x)\, p(y/x)$$

$$= -\sum_x \sum_y p(x,y) \log\left(p(x) \cdot p(y/x)\right)$$

$$= -\sum_x \sum_y p(x,y) \log p(x) \; - \sum_x \sum_y p(x,y) \log p(y/x)$$

$$-\sum_x \left(\sum_y p(x,y)\right) \log p(x) \qquad H(Y/X)$$

$$p(x)$$

$$= -\sum_x p(x) \log p(x) + H(Y/X) = H(X) + H(Y/X)$$

Eg. Let (X, Y) have the following joint distribution.

X \ Y	0	1
0	1/3	1/3
1	0	1/3

$H(X|Y)$; $H(Y|X)$; $H(X,Y)$

$$H(X|Y) = \sum_y p(y)\, H(X|Y=y)$$

~~$p(X|Y=0)$~~ =

$$p(X=0|Y=0) = \frac{p(X=0, Y=0)}{p(Y=0)}$$

	X = 0	X = 1
$p(X|Y=0)$	1	0
$p(X|Y=1)$	1/2	1/2

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

$H(X|Y=1) = 1$

0	1	
2/3	1/3	→ X
1/3	2/3	→ Y

marginal distribution

$H(X|Y) = \frac{1}{3} \times 0 + \frac{2}{3} \times 1$

$= 2/3$ bits

$$H(Y|X) = \sum_x p(x)\, H(Y|X=x)$$

	Y = 0	Y = 1
$p(Y|X=0)$	1/2	1/2
$p(Y|X=1)$	0	1

$H(Y/X=0) = 1$

$H(Y/X=1) = 0$

$H(Y/X) = \frac{2}{3} \times 1 + \frac{1}{3} \times 0 = \frac{2}{3}$ bits

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

$= \left(\frac{2}{3} \times 0 + \frac{1}{3} \times 1\right) + \frac{2}{3}$

$= \frac{1}{3} + \frac{2}{3} = 1$ bits

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

1	2	3	4	
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\rightarrow X$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\rightarrow Y$

marginal distribution.

1 2 3 4 $\rightarrow X$

$P(X/Y=1)$

$(-\infty, x]$ → Half intervals in $B(R)$

$A_i = (-\infty, x - \frac{1}{i}]$ for $i = 1, 2, \ldots \infty$

$\bigcup_{i=1}^{\infty} A_i = (-\infty, x)$ (x is not included)

When $X \to g(X)$

where ~~X is~~ $g(X)$ is a one-one function,

then the entropy of X and $g(X)$ is same

Eg. 1) $X, Y = g(X) = 2^X + X + 1$

$\therefore H(X) = H(Y)$

2.) $X, Y = X^2$

$P_1 \to 1$

$P_2 \quad 1 \to 1$

$-P_1 \log P_1 - P_2 \log P_2 \geq -(P_1+P_2)\log(P_1+P_2)$

Proof: $P_1 + P_2 \geq P_1$, $P_1 + P_2 \geq P_2$

$\Rightarrow \log(P_1+P_2) \geq \log P_1$, $\log(P_1+P_2) \geq \log P_2$

$\Rightarrow P_1 \log(P_1+P_2) \geq P_1 \log P_1$, $P_2 \log(P_1+P_2) \geq P_2 \log P_2$

$\Rightarrow (P_1+P_2)\log(P_1+P_2) \geq P_1 \log P_1 + P_2 \log P_2$

$\Rightarrow -(P_1+P_2)\log(P_1+P_2) \leq -P_1 \log P_1 - P_2 \log P_2$

Entropy

$H(X) = E[g(X)]$

$g(X) = \log_2 \frac{1}{P_X(x)}$

R.V under consideration are discrete R.V

X	0	1	2	3	4
	$P_X(0)$	$P_X(1)$	$P_X(2)$	$P_X(3)$	$P_X(4)$

$P_X(x)$ itself is a R.V.

$P_X(x)$ is a function of x & hence a R.V

$$g(x) = \log_2 \frac{1}{P_x(x)}$$

avg length = 2 bits / symbol

Eg.1.) source → 1 2 3 4 ⇒ 1→00, 2→01, 3→10, 4→11

1/4 1/4 1/4 1/4

1 2 3 3.

2) Prefix free condition

1→0, 2→10, 3→110, 4→111

source: 1 2 3 4 ~~Pbee~~ Prefix free condition

1/2 1/4 1/8 1/8 - one codeword

1 2 3 3. the codeword for one source (1,2,3,4) donot coincide with any other.

1 2 3 4 2 1

0 10 110 111 10 0. → above source can be obtained from this code when the digits for each source (1,2,3,4), are unique

Eg 0 → 1

01 → X

~~Or~~ after 0 is not ~~posth~~ possible.

10 → 2, 101(x), 011(x), ..

each one is unique, so we are able to identify/ derive source from code.

Average length = $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8}$

$= 7/4 < 2$

~~Entropy~~ Average length = 2 bits/symbol for uniform distribution.

Entropy = measure of uncertainty in the source or amount of info in the source.

Joint Entropy (for joint pmf)

$$H(X,Y) = \sum_{x \in X} \sum_{y \in Y} P_{X,Y}(x,y) \log \frac{1}{P(x,y)}$$

$$H(X/Y) = \sum_{x \in X} \sum_{y \in Y} P_{X,Y}(x,y) \log \frac{1}{P_{X/Y}(x/y)}$$

Conditional entropy

$P_{X,Y}$

Y \ X	1	2	3	4	
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	$\rightarrow P(Y=1)=\frac{1}{4}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$	$\rightarrow P(Y=2)=\frac{1}{4}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\rightarrow P(Y=3)=\frac{1}{4}$
4	$\frac{1}{4}$	0	0	0	$\rightarrow P(Y=4)=\frac{1}{4}$
	$P(X=1)=\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	

$$H(X/Y) = \sum_{y} P_Y(y) H(X/Y=y)$$

$$P_{X/Y=y}(x/y) = \sum_{x} P_{X/Y=y}(x/y) = 1$$

$$P_{X/Y=1} = \frac{P_{X,Y}(x,1)}{P_Y(1)} \Rightarrow \frac{\frac{1}{8}}{\frac{1}{4}} \quad \frac{\frac{1}{16}}{\frac{1}{4}} \quad \frac{\frac{1}{32}}{\frac{1}{4}} \quad \frac{\frac{1}{32}}{\frac{1}{4}} \qquad (Y=1)$$

$$\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{8}$$

$H(X)$ = amount of uncertainity in x or amount of information in x.

$H(X/Y)$ = Amount of uncertainity in x even after

observing or amount of info in x even after observing y.

$H(X/Y) \leq H(X)$ → amount of uncertainity of info in x

↓ amount of uncertainity of info in x given y

$H(X/Y) = H(X)$ (when x, y are independent)

Suppose if $X = g(Y)$

$H(g(Y)/Y) = 0$ as if y is known, then $g(Y)$ is also known

Mutual Information between two r.vs X & Y

$P_{X,Y}$ (joint distribution)

~~I(x)~~

$I(X;Y) = H(X) - H(X/Y)$

- ; → semicolon
- $H(X)$ → amount of uncertainity in X
- $H(X/Y)$ → amount of uncertainity in X even after observing Y.
- $I(X;Y)$ → amount of information Y is giving about X.

Chain Rule for Entropy

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

$= H(Y) + H(X/Y)$

$\Rightarrow H(X) - H(X/Y) = H(Y) - H(Y/X)$

$\boxed{I(X;Y) = I(Y;X)}$

Joint pmf = product of marginal pmf and conditional pmf.

$P_{X,Y}(x,y) = P_X(x) \cdot P_{Y/X}(y/x)$

Products in probability get converted to sums

in entropy as we are taking log.

In communication,

x (transmitted), Y (receiver).

Relative Entropy

between 2 prob's.

is # proxy for distance between two prob's:

$P(x), q(x), \quad x \in X.$

$$D(p \| q) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

contribution to sum $\rightarrow$ $= 0$ if $p(x) = 0, q(x) \neq 0$

$p(x) \neq 0, q(x) = 0 \longrightarrow +\infty$

$p(x) = 0, q(x) = 0 \longrightarrow 0.$

$$D(p \| q) \neq D(q \| p)$$

Relative entropy between joint prob's of x & y & product of marginals of x & y

$$I(X;Y) = D(P_{xy} \| P_x P_y)$$

relative entropy gives how far are the r.v's from being independent

Proof: $I(X;Y) = H(X) - H(X/Y)$

$$D(P_{xy} \| P_x P_y) = \sum_{x \in X, y \in Y} P_{xy}(x,y) \log \left(\frac{P_{xy}(x,y)}{P_x(x) P_y(y)} \right)$$

$$= \sum_{x,y} P_{x,y}(x,y) \log \frac{P_{x/y}(x/y)}{P_x(x)}$$

$$= \sum_{x,y} P_{x,y}(x,y) \log \frac{1}{P_x(x)}$$

$$= \sum_{x} \left(\sum_{y} P_{x,y}(x,y) \log \frac{1}{P_x(x)} \right)$$

$$P_x(x)$$

$$= \sum_{x} P_x(x) \log \frac{1}{P_x(x)}$$

Prove that: $D(p\|q) \geq 0$.

Hint: use the fact that $\forall z \in \mathbb{R}, \ln z \leq z-1$

$$I(X;Y) = H(X) - H(X/Y)$$

$$I(X;Y) = D(p\|q)$$

$$D(p\|q) \geq 0 \Rightarrow \boxed{H(X) \geq H(X/Y)}$$

$$D(p\|q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = -\sum_{x} p(x) \log \frac{q(x)}{p(x)}$$

$$D(p\|q) \geq \sum_{x} p(x) \left(\frac{q(x)}{p(x)} - 1 \right)$$

$$D(p\|q) \geq 0.$$

Quiz-1

1.) (i) $x^2(t) \xleftrightarrow{\mathcal{F}} X(j\omega) * X(j\omega)$

$x_1(t) \cdot x_2(t) \xleftrightarrow{\mathcal{F}} 2\pi X_1(j\omega) * X_2(j\omega)$

$$\int_{-\infty}^{\infty} x(\tau) y(t-\tau) d\tau$$

Hint:

$x(t) * y(t) \xleftrightarrow{\mathcal{F}} X(j\omega) Y(j\omega)$

$x(t) \cdot y(t) \xleftrightarrow{\mathcal{F}} c \cdot X(j\omega) * Y(j\omega)$

Inverse transform of $2\pi X_1(j\omega) * X_2(j\omega)$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} 2\pi X_1(j\omega) * X_2(j\omega)\, e^{j\omega t}\, d\omega$$

$$= \int_{w=-\infty}^{\infty}\left(\int_{\tilde{w}=\infty}^{\infty} X_1(j\tilde{w})\, X_2(j\alpha(w-\tilde{w}))\, d\tilde{w}\right)$$

$=$

(ii). $x(t) + x(t-1) \xrightarrow{\mathcal{F}} X(j\omega) + X(j\omega)\, e^{-j\omega\cdot 1}$

$= X(j\omega)(1 + e^{-j\omega})$

Nyquist rate $= \omega_0$.

2) $\mathcal{F}_1$ and $\mathcal{F}_2$ → event space

Is $\mathcal{F}_1 \cup \mathcal{F}_2$ an event space? → No.

$\mathcal{F}_1 = \{\phi, \Omega, A, A^c\}$. with A, B — disjoint

$\mathcal{F}_2 = \{\phi, \Omega, B, B^c\}$. $A \cap B \neq \phi, \Omega$

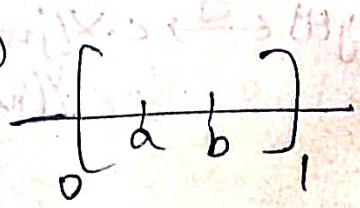
$A \cup B \neq \Omega$

$\mathcal{F}_1 \cup \mathcal{F}_2 = \{\phi, \Omega, A, A^c, B, B^c\}$

Missing $A \cup B$.

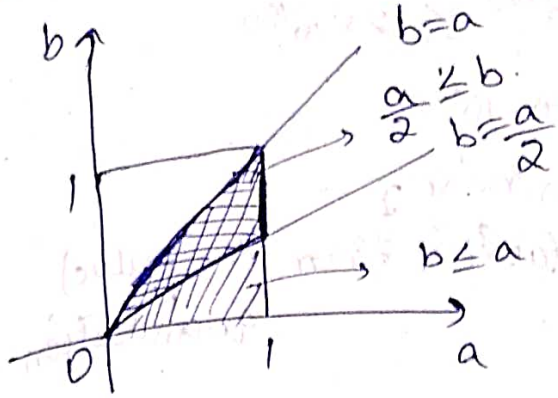
$\mathcal{F}_1 \cap \mathcal{F}_2$ → Yes, it is an event space

3.)



what is $P\left(1 \le \frac{a}{b} \le 2\right)$

$= P(b \le a \le 2b)$



Area = shaded region
$= \frac{1}{4}$.

When there are 2 random variables independent of each other, then their joint density will be the product of the two densities.

$F_a(a) = a$, $F_b(b) = b$. ⟶ CDF

$f_{ab}(a,b) = f_a(a) \times f_b(b)$ ⟶ PDF

$= 1.1 = 1$, $0 \le a, b \le 1$

0, otherwise

Midsem

Q1.) $y(t) = x_1(t) . x_2(t)$

$y(t) \xleftrightarrow{\mathcal{F}} 2\pi X_1(j\omega) * X_2(j\omega)$

Bandwidth = sum of bandwidths of both x_1 and x_2

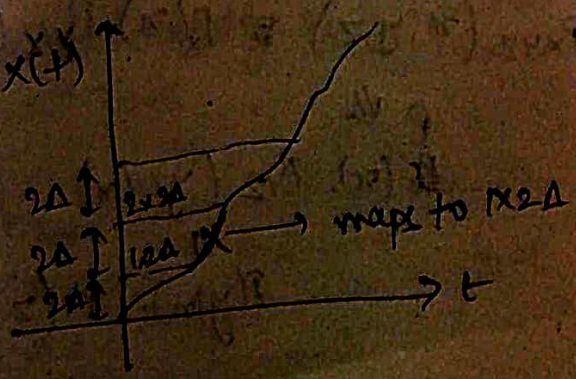
= Bandwidth $x_1(j\omega)$ + BW of $x_2(j\omega)$

$= 1000\pi + 2000\pi$

$= 3000\pi$

So Nyquist rate $= 2 \times 3000\pi = 6000\pi$

Q3.) Step size $= 2\Delta$.



Eg. $10^{-120} \leq 10^{-120} + 10^{-122} \leq 2 \times 10^{-120}$

↓ maps to 10^{-120}

10^{-122} → quantisation noise ?
(how much far it is from the value)

quantisation.

P.D.F

$f_e(e) = \frac{1}{2\Delta}$ ⇒ $\int_{-\infty}^{\infty} f_e(e)\,de = 1$

$= \int_{-\Delta}^{\Delta} f_e(e)\,de = \int_{-\Delta}^{\Delta} x \frac{1}{2\Delta}\,de = 1$

X → uniform random variable ([a, b])

$$f_x(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

Given: e ~ uniform random variable ([−Δ, Δ]):

$$f_e(e) = \begin{cases} \frac{1}{2\Delta}, & -\Delta \leq e \leq \Delta \\ 0, & \text{otherwise} \end{cases}$$

$E[e] = \int_{-\infty}^{\infty} e \cdot f_e(e)\,de = \int_{-\Delta}^{\Delta} e \frac{1}{2\Delta}\,de$

$= \frac{1}{2\Delta}\left(\frac{e^2}{2}\right)\Big|_{-\Delta}^{\Delta} = \frac{1}{2\Delta} \times \frac{1}{2}(\Delta)^2 - (\Delta)^2 = 0$

Variance = $E[x^2] - (E[x])^2$

4.) $P_{xyz}(x, y, z) = P(x) \cdot P_{y/x}(y/x) \cdot P_{z/y}(z/y)$

$\Downarrow$

$P_x(x) \cdot P_{yz}(yz/x)$

$\Downarrow$

$P(y/x) \cdot P(z/y, x)$

$= P_X(x) \cdot P_{Y/X}(y/x) \cdot P_{Z/Y,X}(z/y,x)$

(chain rule)

$= P_X(x) \cdot P_{Y/X}(y/x)\, P_{Z/Y}(z/y)$ → given in Q

$\Rightarrow P_{Z/Y,X}(z/y,x) = P_{Z/Y}(z/y)$

$\Rightarrow \boxed{P(z/y) = P(z/y,x)}$ [$Z \perp\!\!\!\perp X$ (conditioned on Y)]

Z is independent of X conditioned on Y.

$P(yz) = P(y) \cdot P(z/y)$

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

$\Rightarrow$ Usually $H(X,Y,Z) = H(X) + H(Y,Z/X)$

$= H(X) + H(Y/X) + H(Z/Y,X)$

$P_{ZXYZ} = P_X(x) \cdot P_{Y/X}(y/x) \cdot P_{Z/Y}(z/y)$

$\boxed{P(z/y) = P(z/y,x)}$ $\forall z,y,x$

$H(X,Y,Z) = H(X) + H(Y,Z/X)$

$= H(X) + H(Y/X) + H(Z/Y,X)$

Markov condition.

$\Rightarrow Z \perp\!\!\!\perp X$ conditioned on Y

$H(Z/Y)$

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

2.) Tossing X coins, such that exactly r heads have appeared until the x^{th} toss.

(a). $P(X=x) = {}^{n-1}C_{r-1}\,(p)^r\,(1-p)^{n-r}$.

$= p \cdot \binom{n-1}{r-1} \cdot (1-p)^{(n-1)-(r-1)} \cdot p^{r-1}$

$= p^r \cdot \binom{n-1}{r-1} \cdot (1-p)^{n-r}$

r negative Binomial

Suppose $r=1$, $P(X=n) = p(1-p)^{n-1}$

[X is a geometric variable of p]

$$E[X] = \frac{1}{p}$$

In general for $r=r$. $(r>1)$

$$\Rightarrow \boxed{E[X] = \frac{r}{p}}$$

Source Coding (compressing source to its entropy)

Operational Significance - practically what does the quantity indicate.

Let us have an information source → X

X takes values from $\mathcal{X}$.

PMF of X is denoted by P_X.

Source code is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$.

$\mathcal{D}^*$ - set of all binary strings.

$= \{ 0, 1, 00, 10, 01, 11, 000, 100, \ldots \}$

all binary strings

(countably infinite terms).

$C : \mathcal{X} \longrightarrow \mathcal{D}^*$

Source code 1

$1 \longrightarrow 00$	$C(1) = 00$
$2 \longrightarrow 01$	$C(2) = 01$
$3 \longrightarrow 10$	$C(3) = 10$
$4 \longrightarrow 11$	$C(4) = 11$

Source Code 2

$1 \longrightarrow 0$	$C(1) = 0$
$2 \longrightarrow 10$	$C(2) = 10$
$3 \longrightarrow 110$	$C(3) = 110$
$4 \longrightarrow 111$	$C(4) = 111$

Examples of source code

1.) ASCII - character code for English text

7-bit code.

33 control characters
95 printing characters. } - 128 characters.

(fixed length code)
equal no. of bits for all characters.

variable length code → the no. of bits are not same for all elements in source code

Morse code

characters more frequently used should be given lesser number of bits,
less frequently used characters should be assigned more no. of bits.

Morse code using Dots & Dashes → short pulse
→ long pulse.

A .— → more frequently
Z ——.. → least frequently
I .. → most frequently
Q ——.— → less frequently

(pulse width of dash is 3 times that of dot)

C: X → D*
- X → symbols of the source
- D* → codewords.
  - . → more frequent
  - — → less frequent
  - o, oo → they are different.

Prefix-free source code

A code C is said to be prefix free if no codeword is the prefix of any other codeword.

C(1) = 1
C(2) = 10
C(3) = 01
C(4) = 110

C(2) = 10, C(4) = 110
C(1)
so the codeword of C(1) is a prefix of codeword of C(2), C(4)

not prefix free code

Why is prefix free ~~code~~ property important?

Important for instant decodability

as you are parsing the coded sequence, as soon as the codeword ends, you can decode

0 1 0 1 1 0 1 0 1 1 1 0 0 1 0 0

1 2

1 2 3 2 2 1

1 10

(self functionality)

$X \quad D^*$

$x \longrightarrow C(x) \qquad l(x)$ = length of codeword assigned to symbol x.

Average length of a source code C is denoted by $L(C)$

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

Eg-1) X. $\quad P(1) = P(2) = P(3) = P(4) = 1/4$

C_1 $\quad 1 \to 00,\ 2 \to 01,\ 3 \to 10,\ 4 \to 11$

C_2 $\quad 1 \to 0,\ 2 \to 10,\ 3 \to 110,\ 4 \to 111$

$L(C) = H(X)$

$H(X) = 2$ bits/symbol

$L(C_1) = 2$ bits/symbol

$L(C_2) = \frac{1+2+3+3}{4} = 2.25$ bits/symbol

this code is better for this source.

*2.) $X \quad P(1) = 1/2,\ P(2) = 1/4,\ P(3) = 1/8,\ P(4) = 1/8$

$C_1 \to 1 \to 00,\ 2 \to 01,\ 3 \to 10,\ 4 \to 11$

$C_2 \quad 1 \to 0,\ 2 \to 10,\ 3 \to 110,\ 4 \to 111$

$L(C_1) = 2$ bits/symbol ; ~~$L(C_2) =$~~

$$L(C_2) = 1\times\frac{1}{2} + 2\times\frac{1}{4} + 3\times\frac{1}{8} + 3\times\frac{1}{8}$$

$$= 1.75 \text{ bits/symbol}$$

$$H(X) = 1.75 \text{ bits/symbol}$$

$$L(C_2) = H(X)$$

Theorem: If a code C is prefix-free for a source X, then

$$\boxed{L(C) \geq H(X)}$$

Kraft's Inequality: If a code is prefix-free, then $\{l(x)\}$ satisfy the following condition.

$$\boxed{\sum_{x\in\mathcal{X}} 2^{-l(x)} \leq 1}$$

Prove the theorem using the following 2 results

1.) Kraft's Inequality

2) $D(p\|q) \geq 0$

Proof: $D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$

Taking $q(x) = \dfrac{2^{-l(x)}}{\sum_{x\in\mathcal{X}} 2^{-l(x)}}$

$$D\left[p(x)\,\|\,\frac{2^{-l(x)}}{\sum_{x\in\mathcal{X}} 2^{-l(x)}}\right] \geq 0$$

$$\Rightarrow \sum p(x)\log\left[\frac{p(x)\cdot\sum 2^{-l(x)}}{2^{-l(x)}}\right] \geq 0.$$

$$\sum p(x)\log p(x) + \sum p(x)\log\left(\sum 2^{-l(x)}\right) \geq$$

$\sum p(x)\log p(x)$: $-H(X)$; $\sum p(x)\log\left(\sum 2^{-l(x)}\right)$: ≤ 1 ; $\sum p(x)\log\left(2^{-l(x)}\right)$: ≤ 0 ; $\sum p(x)\log\left(2^{-l(x)}\right)$: $\leq -L(C)$

$$\Rightarrow -H(X) + (-L) \geq (-L(C))$$

$$\Rightarrow H(X) + K \geq L(C)$$

$$\Rightarrow \boxed{H(X) \leq L(C)}$$

Assuming,

$$L(C) - H(X) \geq 0$$

$$\sum l(x)p(x) + \sum p(x)\log(p(x)) \geq 0$$

$$\Rightarrow +\sum p(x)\log\left(\frac{1}{2^{-l(x)}}\right) + \sum p(x)\log p(x) \geq 0.$$

$$\Rightarrow \sum p(x)\log\left(\frac{p(x)}{2^{-l(x)}}\right) \geq 0$$

$\Rightarrow$ From $D(p\|q) \geq 0$, we know that

$$\sum p(x)\log\left(\frac{p(x)}{\left(\frac{2^{-l(x)}}{\sum 2^{-l(x)}}\right)}\right) \geq 0$$

$$\Rightarrow \sum_x p(x)\log\frac{p(x)}{2^{-l(x)}} + \sum_x p(x)\log\left(\sum 2^{-l(x)}\right) \geq 0.$$

$$L(C) \geq H(X)$$

x	1	2	3	4

$P(1) = 2/3$, $P(2) = \frac{1}{6}$, $P(3) = \frac{1}{12}$, $P(4) = \frac{1}{12}$

There are several sources where ~~L(C)~~ $L(C) \neq H(X)$ or ~~very~~ $L(C)$ is not very close to $H(X)$

Only for distributions with powers of 2, H.

$$L(C) = H(X)$$

Designing an optimal prefix-free source code for a general source is equivalent to solving this optimization problem.

$$\min \sum_{x \in \mathcal{X}} l(x) P_X(x)$$
$$\{l(x)\}$$

such that

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

To solve the optimization problem, a technique known as Huffman Coding is used.

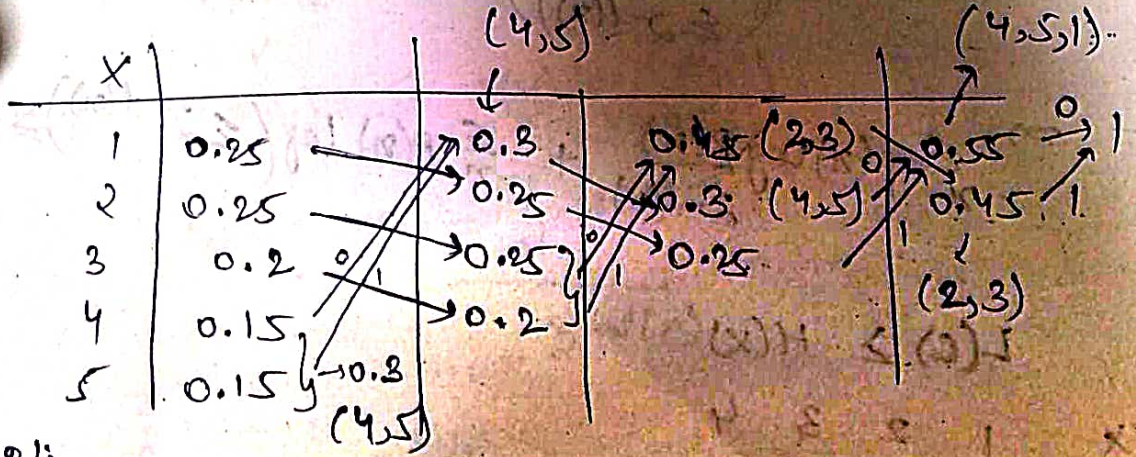
↓

(JPEG).

Example of Huffman Coding → algorithm can be used for any source.

$P(1) = 0.25$, $P(2) = 0.25$, $P(3) = 0.2$

$P(4) = 0.15$, $P(5) = 0.15$



Step 1: Combine the least 2 probabilities into a super symbol

Step 2: Keep repeating the same process until you get only a single super symbol

(Binary code, so we are combining 2 symbols)

⇒ 1 → 01
2 → 10
3 → 11
4 → 000
5 → 001

(the 0's and 1's are taken from behind (i.e. from 1 to 0.55, 0.45 to ...) in reverse direction

If a code is prefixfree, it satisfies Kraft's inequality.

$$\sum_{x \in X} 2^{-l(x)} \leq 1$$

If a code is prefixfree, then $L(C) \geq H(X)$

In general, we should solve the following optimization problem.

$$\min_{\{l(x)\}} \sum_{x \in X} l(x) P_X(x)$$

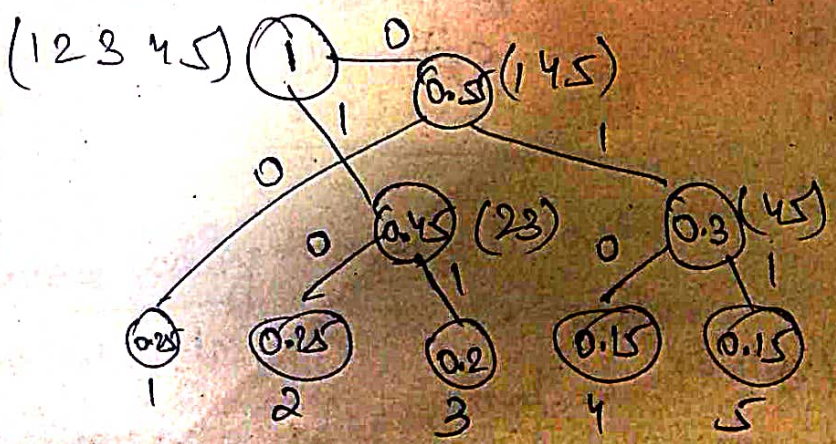
$$\sum_{x \in X} 2^{-l(x)} \leq 1$$

↙

Huffman Algorithm / Huffman code

Eg.

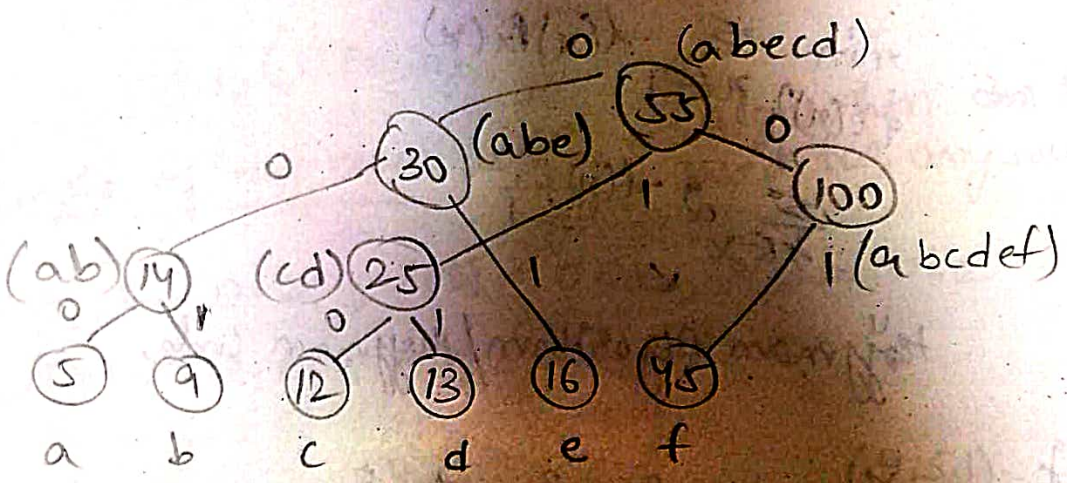
X	1	2	3	4	5
$P_X(x)$	0.25	0.25	0.2	0.15	0.15



1 → 00
2 → 10
3 → 11
4 → 010
5 → 011

Eg. 2.

characters	frequency
a	5
b	9
c	12
d	13
e	16
f	45



a → 0000
b → 0001
c → 010
d → 011
e → 001
f → 1

for any source, Huffman code is optimal

$L(C) \geq H(X)$

$L(C) \leq H(X)+1$

Any source code ~~with~~ which follows Kraft's inequality, can be written in the form of prefix free code.

$$l(x) = \left\lceil \log_2 \frac{1}{P_x(x)} \right\rceil$$

$$\Rightarrow \sum_{x \in \mathcal{X}} 2^{-\left\lceil \log_2 \left(\frac{1}{P_x(x)}\right) \right\rceil} \leq \sum_{x \in X} 2^{-\log_2 \left(\frac{1}{P_x(x)}\right)}$$

$$\Rightarrow \leq a$$

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

$$\Rightarrow \sum_{x \in \mathcal{X}} P(x) \leq 1 \Rightarrow \text{sum of all probabilities} = 1$$

$$\Rightarrow \boxed{\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1}$$

For the choice of lengths $l(x) = \left\lceil \log_2 \frac{1}{P_x(x)} \right\rceil$

$$L(C) = \sum_{x \in X} P_x(x) \left\lceil \log_2 \frac{1}{P_x(x)} \right\rceil$$

$$< \sum_{x \in X} P_x(x) \left(\log_2 \frac{1}{P_x(x)} + 1 \right)$$

$$< H(x) + 1$$

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

Source coding theorem: A source can be compressed arbitarily close to its Entropy.

IID → independent identically distributed

$H(X) + \epsilon$

IID r.vs $X_1, X_2, \cdots X_n \sim P_X(x)$

$P_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} P_{X_i}(x_i)$

Distribution is similar to $P_X(x)$ as the r.vs are identically distributed.

$$H(X_1, X_2, \cdots X_n) = \sum_{i=1}^{n} H(X_i) = nH(X)$$

A source code can be compressed arbitarily close to the entropy. Designing a source code for some n.

$$nH(X) \leq L(C) < nH(X) + 1$$

$$H(X) \leq \frac{L(C)}{n} \leq H(X) + \frac{1}{n}$$

by varying n, the average length can be made as close to $H(X)$

$\frac{L(C)}{n}$ → average length per symbol / average length per instance per symbol.

If n=2 (from previous eg.)

(1,1) → $(0.25)^2$ → 0000 (appending both source codes)

(1,2) → (0.25)(0.25) → 0010

⋮

(1,5) → (0.25)(0.15) → 0011

⋮

(5,5) → $(0.15)^2$

25

1 → 00

2 → 10

3 → 11

Show the following inequality.

$$\frac{L_2}{2} \leq L_1$$

L_1 = optimal length when $n=1$

L_2 = optimal length when $n=2$.

$$L_1 = \sum_{x \in X} p(x) \cdot \log l(x)$$

$$\frac{L_2}{2} \leq H(X) + \frac{1}{2}$$

$$L_1 \leq H(X) + 1$$

$$H(X) + \frac{1}{2} \leq H(X) + 1$$

$$\Rightarrow \boxed{\frac{L_2}{2} \leq L_1}$$

$$L_2 = \sum_{x \in X} p_2(x) \cdot l_2(x)$$

$$\downarrow$$

$$2 l_1(x)$$

$$\Rightarrow \frac{L_2}{2} = \sum_{x \in X} p_2(x) l_1(x)$$

$$\Rightarrow \sum_{x \in X} p_2(x) l_1(x) \leq \sum_{x \in X} p_1(x) l_1(x)$$

$$\Rightarrow \boxed{\frac{L_2}{2} \leq L_1}$$

$$\frac{L_3}{3} \leq \frac{L_2}{2} \rightarrow \text{prove.}$$

$$\frac{L_3}{3} \leq L_1 \text{ and } L_3 \leq L_1 + L_2 \Rightarrow \frac{L_3}{3} \leq \frac{L_1}{3} + \frac{L_2}{3}$$

$$\Rightarrow \frac{L_3}{3} \leq \frac{L_2}{3} \leq \frac{L_2}{2}.$$

$$L_1 \quad \frac{L_2}{2} \leq L_1$$

$$\Rightarrow \frac{3L_2}{2} \leq L_1 + L_2 \Rightarrow L_3 \leq \frac{3L_2}{2} \Rightarrow \boxed{\frac{L_3}{3} \leq \frac{L_2}{2}}$$

Lempel-Ziv Algorithm (LZ-77, LZ-78)

(core of .zip, .gzip)

uses Dictionary based method.

this method does not even assume source probability.

LZ Algorithm

Eg. String → CDDCDDCDDDCCDCDD

1.) Parse the string from left to right

① → C ⇒ if nothing is there in the dictionary (null) then you add it to the dictionary.

(0, C) ② → D ⇒ (0, D)

③ → D ⇒ (1, 1, 1)

- 1st 1 → which position from the current position is the string repeated
- 2nd 1 → repetition occ so it is 1
- 3rd 1 → length of the match (length of the no of times repeated)

Window length = 4

i.e. position − curr ≤ 4

(C D D)

position ← ③ → curr

from curr position repeated

position curr → diff = 3

CDDCDDCDD

whole string is repeated twice

so size of the repeated string becomes 6.

④ → C

(1, 3, 6)

CDDCDDCDD → over

now CDDCDDCDDDC

position ↑ curr

⇒ (1, 4, 2) → (DC)

C C D C D D

position, curr.

(1, 1, 1)

D C C D C D D

curr

position

(1, 3, 2)

D C C D C D D

curr

(1, 2, 2)

(1, 4, 2)

DC is repeating (we are going back by 4 steps)

window = 4

can move back max 4 steps, i.e. position can be max 4.

(1, 3, 6)

C repeating

3 positions before

we are seeing if CDD (before C) is repeating, its repeating twice after c, so length of string = 6

To get back,

C D D C D D C D D C D D C C D C D D

(0, C) so C unrepeated C

(0, D) so D unrepeated CD

(1, 1, 1) so repeated, 1 step back, one letter → to D

repeating which is D, so CDD

(1, 3, 6) so repeated, 3 steps back, 6 repeating, so CDDCDD...

Tut

2.) $(P(x) : x \in \mathcal{X})$

Find $(l(x) : x \in \mathcal{X})$ such that $\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$

we know that $\sum_{x \in \mathcal{X}} P(x) = 1$

Come up with $(l(x) : x \in \mathcal{X})$ so that $\Rightarrow$ a prefix-code with these lengths.

$$l(x) \triangleq \left\lceil \log_2 \frac{1}{P(x)} \right\rceil \geq \log_2 \frac{1}{P(x)}$$

$\left\lceil \log \frac{1}{P(x)} + 1 \right\rceil$

satisfies. $\sum 2^{l(x)} = \sum_{x \in \mathcal{X}} 2^{-\left\lceil \log_2 \frac{1}{P(x)} \right\rceil}$

$$\leq \sum_{x \in \mathcal{X}} 2^{-\log_2 \frac{1}{P(x)}} = \sum_{x \in \mathcal{X}} P(x) = 1$$

3.) $E[L] = \sum_{l(x)} P(l(x))\, l(x)$

$$= \sum_{x \in \mathcal{X}} P(x) l(x)$$

$$= \sum_{x \in \mathcal{X}} P(x) \left\lceil \log \frac{1}{2 P(x)} \right\rceil$$

$$\leq \sum_{x \in \mathcal{X}} P(x) \left(\log_2 \frac{1}{P(x)} + 1 \right)$$

$$= H(X) + \sum_{x \in \mathcal{X}} P(x) = H(X) + 1$$

$Y = g(X)$

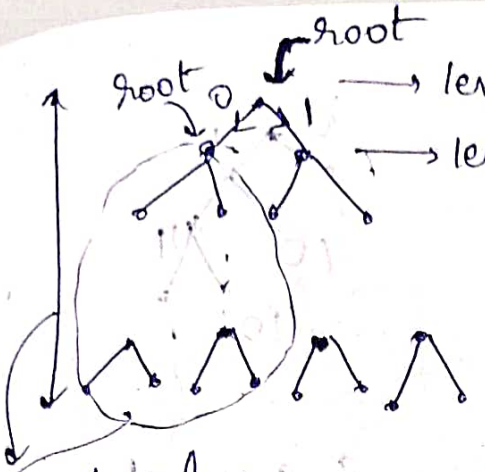
$E[Y] = \sum_y P(y) \cdot y$

$= \sum_x g(x) \cdot P(x)$

as $P(x) = P(y)$ (because if we x is known, then y is also known)

$$H(X) \leq E[L] \leq H(X) + 1$$

↑ corresponds to prefix code



→ level 0 has $2^0 = 1$ node

→ level 1 has $2^1 = 2$ nodes

→ level i has 2^i nodes.

depth $= l_{max}$

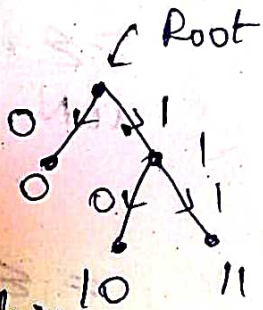
$= \max_{x \in \mathcal{X}} l(x)$

Total nodes $= \sum_{i=0}^{l_{max}} 2^i$

$= 2^{l_{max}+1} - 1$

total nodes $= 2^{l_{max}} - 1$

$l_{min} = \min_{x \in \mathcal{X}} l(x) = 1$



no. of nodes dropped $= 2^{l_{max} - l_{min} + 1}$

(we are dropping all the descendends)

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

$$2^{l_{max}+1} \sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 2^{l_{max}+1}$$

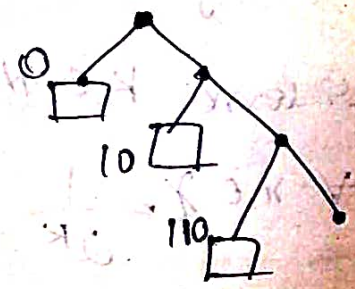
$$\Rightarrow \sum_{x \in \mathcal{X}} 2^{l_{max} - l(x) + 1} \leq 2^{l_{max}+1}$$

(total no. of nodes dropped in our procedure)

we've shown

Kraft's inequality $\Rightarrow$ Prefix-free code

$\Leftarrow$

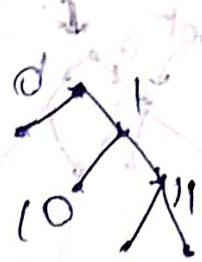


no assigned code is assigned.

4) $p = 0.5$

T, H, TH, TTH, TTTH, ...

0 (1) 01 (2) 001 (3) 0001 (4)



infinitely large tree

$\mathbb{E}[L] = \mathbb{E}[X]$

$X \sim Geo(1/2)$

$= \sum_{l \in \mathbb{N}} l \cdot \frac{1}{2} \times \left(\frac{1}{2}\right)^{l-1}$

head, tail.

$= \sum_{x \in X} x \cdot P(x)$ where $P(x) = \left(\frac{1}{2}\right)^{x-1}\left(\frac{1}{2}\right)$

$\Rightarrow \mathbb{H}(X) = \frac{1}{p} = \frac{1}{1/2} = 2$

$\mathbb{E}[L] = 2.$

$H(X) \leq \mathbb{E}[L] \Rightarrow \boxed{H(X) = \mathbb{E}[L]}$

$H(X) = 2$, $\mathbb{E}[L] = 2$

coding

So the expected scheme we have is optimum coding scheme

5) $X = \{1, 2, 3, \ldots, 2^k\}$ for some $k \in \mathbb{N}$.

$P(x) = \frac{1}{|X|}$, $\forall x \in X = \frac{1}{2^k}$

Goal: Find $x \in X$, using the least no. of questions.

Perform binary search $\Rightarrow$ exactly k questions.

$H(X) = \log\left(\frac{1}{P(x)}\right) = \log|X| = k$

(Since $P(x) = \frac{1}{x}$; $x \sim$ uniformly distributed