Notes on Lecture I : (Entropy & Information Theory) Nilanjana Datta

e.g. Roll a die

r.v. X = number on top face

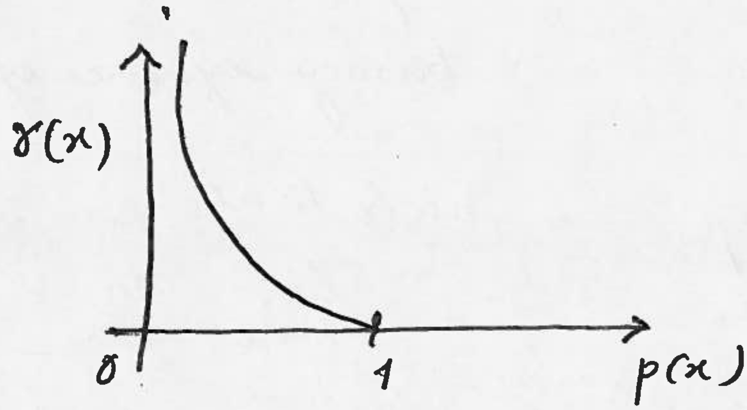
$x \in J = \{1, 2, \cdots 6\}$; $p(x) = 1/6 \ \forall x \in J$

$X \sim p(x)$; $x \in J$

p.m.f.

$P(X=x) = p(x)$

surprisal

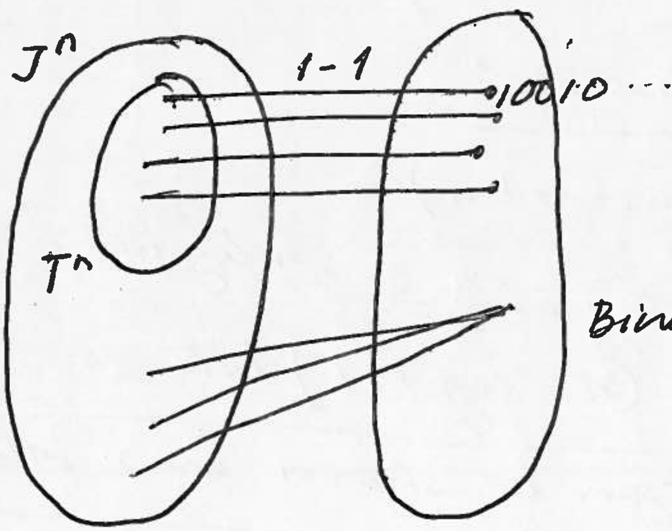


$p(x) = 1 \Rightarrow \gamma(x) = 0$

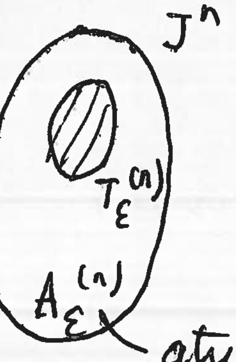
$p(x) \approx 0 \Rightarrow \gamma(x)$ high

$U \sim p(u)$; $u \in J$

$H(U) = -\sum p(u) \log p(u)$



Binary strings of fixed length



atypical

$J^n = T_\varepsilon^{(n)} \cup A_\varepsilon^{(n)}$

$T_\varepsilon^{(n)} \cap A_\varepsilon^{(n)} = \phi$

$\Rightarrow \forall \delta > 0$ n large enough $P(A_\varepsilon^{(n)}) \leq \delta$

$p(\underline{u}) \approx 2^{-nH(U)}$ $\forall \underline{u} \in T_\varepsilon^{(n)}$

$|T_\varepsilon^{(n)}| \approx 2^{nH(U)}$

$P(T_\varepsilon^{(n)}) \approx 1$

Compression scheme of rate $R > 0$

$$\mathcal{E}^n : \underline{u} = (u_1 \cdots u_n) \longmapsto \underline{x} = (x_1 \cdots\cdots x_{m_n})$$

$\underline{u} \in J^n$

$x_i \in \{0, 1\}$.

binary sequence of length m_n.

$\mathcal{E}^n(\underline{u}) = \underline{x}$

message codeword

such that

$$\lim_{n\to\infty} \frac{m_n}{n} = R$$

i.e. $m_n \approx 2^{nR}$

m_n : no. of bits in a codeword

n : no. of uses of the source.

(B) When is this a compression?

$\underline{u} \in J^n$ $\qquad$ $\underline{u} = (u_1 \cdots u_n)$

$\#$ of possible sequences $= |J|^n = 2^{n \log |J|}$

~~##~~ can be stored in $(n \log |J|)$ bits.

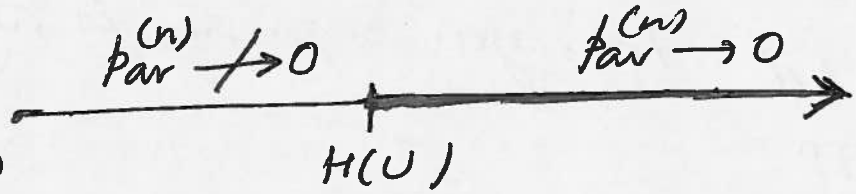
since $\boxed{2^r \text{ messages can be stored in } r \text{ bits}}$

$\therefore$ Compression if $\boxed{n \log |J| > m_n}$

Aim : To make R as small as possible under the requirement $p_{av}^{(n)} \to 0$ as $n \to \infty$

Shannon's Source Coding Theorem

(a) • If $R > H(U)$ $\exists$ a reliable compression scheme of rate R.



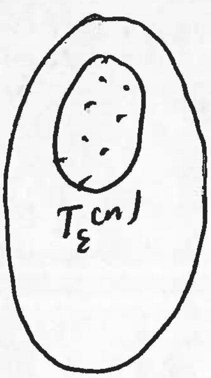
(b) If $R < H(U)$ any compression scheme of rate R is <u>not</u> reliable

<u>Proof of (a)</u> : $R > H(U)$

- Choose $\varepsilon > 0$ s.t. $R > H(U) + \varepsilon$.
- Choose n large enough s.t. Typical sequence Theorem holds.

$$|T_\varepsilon^{(n)}| \le 2^{n(H(U)+\varepsilon)} < 2^{nR}.$$

- Order sequences in $T_\varepsilon^{(n)}$; i.e,
  - give them labels
  - each sequence represented by its labels.
  - no. of bits needed to store these labels is at most nR.



- Inspect the output of n uses of source.

$\underline{u} = (u_1 \cdots u_n)$

(Q) Is it typical — (i) Yes: assign it a unique binary string of length $\lceil nR \rceil + 1$ bits

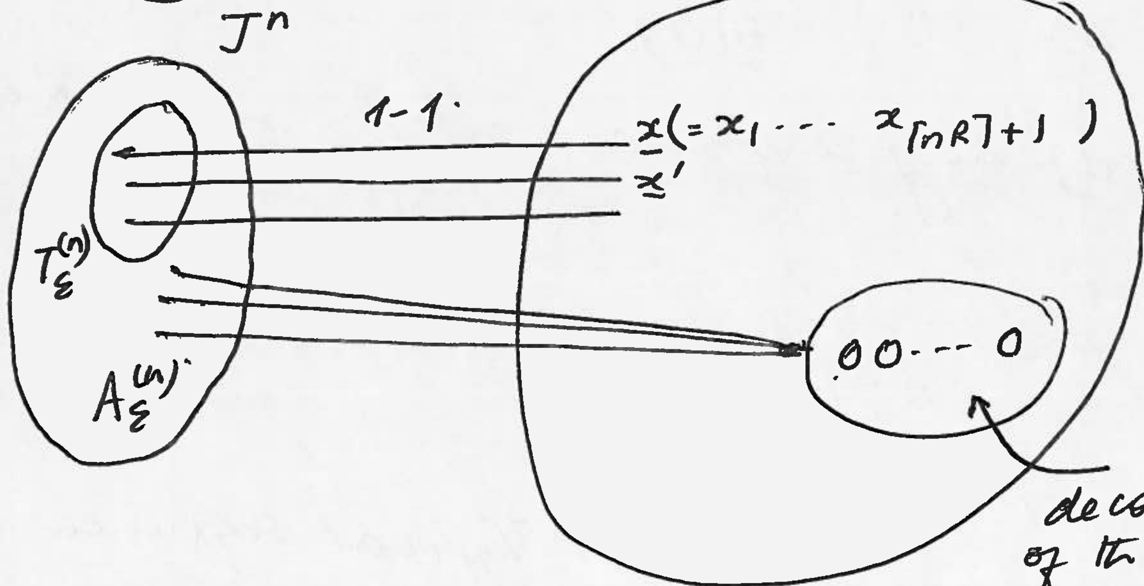
e.g. 1010 . . .

← flag bit ; tells you sequence is typical

(ii) No : map it to a fixed binary string of length $\lceil nR \rceil + 1$

$0\,0 \cdots 0$

flag bit : tells you sequence is atypical.



$\therefore$ information is lost.

But still : $p_{av}^{(n)} \xrightarrow[n\to\infty]{} 0$

Why? since $P(A_\varepsilon^{(n)}) \xrightarrow[n\to\infty]{} 0$.

Details : [let $\underline{u}'$ denote decompressed signal corresponding to $\underline{u}$
For n large enough :

$$p_{av}^{(n)} = \sum_{\underline{u}} p(\underline{u})\, P(\underline{u}' \neq \underline{u})$$

$$= \sum_{\underline{u} \in T_\varepsilon^{(n)}} p(\underline{u}) \underbrace{P(\underline{u}' \neq \underline{u})}_{= 0} + \sum_{\underline{u} \in A_\varepsilon^{(n)}} p(\underline{u}) \underbrace{P(\underline{u}' \neq \underline{u})}_{\leq 1}$$

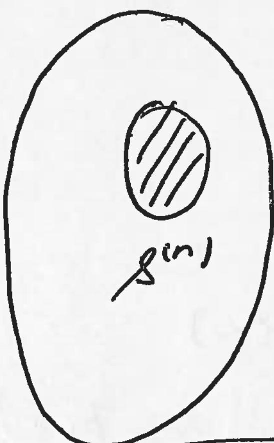
($\because$ 1-1 map for $\underline{u} \in T_\varepsilon^{(n)}$)

$$\therefore\quad p_{av}^{(n)} \le \sum_{\underline{u}\in A_\varepsilon^{(n)}} p(\underline{u}) = P(A_\varepsilon^{(n)}) < \delta$$

$$\therefore\quad \lim_{n\to\infty} p_{av}^{(n)} = 0$$

<u>Rate</u> $\displaystyle \lim_{n\to\infty} \frac{\lceil nR\rceil + 1}{n} = R.$

<u>Converse</u> Any compression scheme of rate $R < H(U)$ is not reliable.



Compression scheme of rate R:

$\Rightarrow$ $\boxed{\text{at most } 2^{nR} \text{ signals assigned } \underline{\text{unique}} \text{ codewords}}$

<u>Lemma</u> $\forall\ \mathcal{S}^{(n)} \subset J^n$ s.t. $|\mathcal{S}^{(n)}| = 2^{nR}$

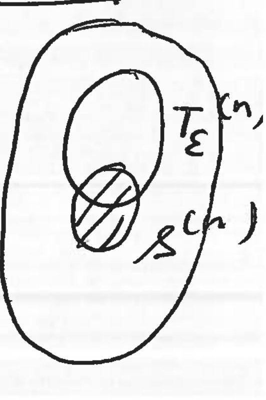
$R < H(U)$ fixed.

$\forall\ \delta > 0$ & n large enough

$$\text{Prob}(\mathcal{S}^{(n)}) < \delta$$

<u>Pf.</u> $\displaystyle P(\mathcal{S}^{(n)}) = \sum_{\underline{u}\in\mathcal{S}^{(n)}} p(\underline{u})$

$$= \sum_{\substack{\underline{u}\in T_\varepsilon^{(n)} \\ \underline{u}\in\mathcal{S}^{(n)}}} p(\underline{u}) + \sum_{\substack{\underline{u}\in A_\varepsilon^{(n)} \\ \underline{u}\in\mathcal{S}^{(n)}}} p(\underline{u})$$



$$= |S^{(n)}| \cdot 2^{-nH(U)} + \delta \qquad \because P(A_\varepsilon^{(n)}) < \delta.$$

$$< 2^{nR - nH(U)} \qquad \because |S^{(n)}| < 2^{nR}.$$

$$= 2^{-n(H(U)-R)} \xrightarrow[n\to\infty]{} 0 \qquad \because R < H(U).$$

<u>Entropies.</u>

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

$$= \sum_x p(x) \Big(-\sum_y p(y|x) \log p(y|x)\Big)$$

$$= -\sum p(x,y) \log p(y|x)$$

$$\because p(x,y) = p(x)\, p(y|x)$$

<u>Chain Rule</u> $\quad H(X,Y) = H(Y|X) + H(X)$

$$RHS = -\sum p(x,y) \log \frac{p(x,y)}{p(x)} - \sum p(x) \log p(x)$$

$$= -\sum_{x,y} p(x,y) \log p(x,y) + \cancel{\sum_{x,y} p(x,y) \log p(x)}$$

$$\cancel{-\sum_x p(x) \log p(x)}$$

$$\because \sum_y p(x,y) = p(x)$$

$$= H(X,Y)$$

$D(p\|q) \geq 0$ $\qquad p = \{p(x)\}$; $q = \{q(x)\}$ $\quad x \in J$

Pf. $D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

let $M = \{x : p(x) > 0\}$.

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

Let A be a r.v. taking values $\frac{q(x)}{p(x)}$ with prob. $p(x)$, $x \in M$

$$\Rightarrow \boxed{-D(p\|q) = \mathbb{E}(\log A)} \text{ — ①}$$

expectation.

$\log u$: concave function of u.

Jensen's inequality

$\forall$ concave function $f(A)$; A : a r.v.

$$\boxed{\mathbb{E}(f(A)) \leq f(\mathbb{E}(A))} \text{ — ②}$$

In ① : $f(A) = \log A$ $\quad$ ∴ using ② in ① :

$$-D(p\|q) \leq \log(\mathbb{E}(A)) = \log\left(\sum_{x \in M} \cancel{p(x)} \frac{q(x)}{\cancel{p(x)}}\right)$$

$$= \log\left(\sum_{x \in M} q(x)\right) \leq \log\left(\underbrace{\sum_{x \in J} q(x)}_{=1}\right)$$

$$= \log 1 = 0$$

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

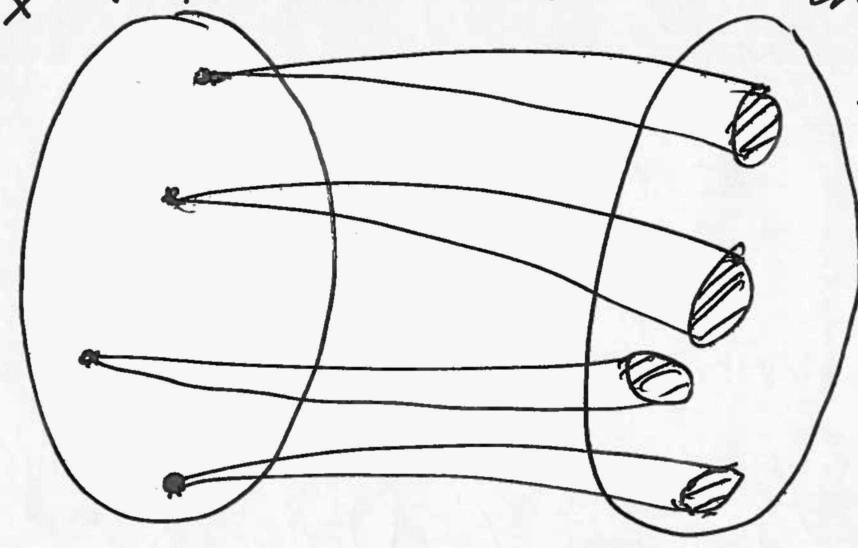
$H(X) \leq \log |J|$ if $X \sim p(x)$ $x \in J$.

Pf: $\underbrace{D(p\|q) \geq 0}$ choose $p = \{p(x)\}$
$q = \{1/|J|\}$ uniform distributia.

$$-H(x) - \underbrace{\sum_x p(x)}_{=1} \log \frac{1}{|J|} \geq 0$$

$$\Rightarrow H(x) \leq \log |J|$$

J_X^n (input sequences for n uses of a memoryless channel)



choose these as code words.

(ii) $\exists$ a set of sequences s.t with high prob. their images do not overlap.

$\therefore$ Correct inference possible.

Shannon's ~~So~~ Noisy Channel Coding Theorem

$$C(\mathcal{N}) = \max_{p(x)} I(X:Y)$$

$$\xrightarrow[\text{input}]{X \sim p(x)} \boxed{\mathcal{N}} \xrightarrow{Y}$$

Intuitive idea behind Shannon's result

- For each x-sequence there are $\approx 2^{nH(Y|X)}$ typical y-sequences. — ⓐ

~~# of typicaly sequ~~

# of typical y-sequences $\approx 2^{nH(U)}$ — ⓑ

[Instead of $p(x)$ we now have conditional probabilities $p(y|x)$

$\therefore$ Relevant entropy $= -\sum p(x)\, p(y|x) \log p(y|x)$
$= H(Y|X)$]

$|T_{\varepsilon,Y}^{(n)}| \equiv$ no. of typical Y-sequences $\approx 2^{nH(Y)}$

Sphere packing:
Divide the set $T_{\varepsilon,Y}^{(n)}$ into disjoint sets of size $\approx 2^{nH(Y|X)}$

$$\therefore \#\text{ of disjoint sets} \approx \frac{2^{nH(Y)}}{2^{nH(Y|X)}} \equiv 2^{nI(X:Y)}$$

$M = \#$ of codewords $\equiv$ no. of disjoint sets

Maximum rate

$$R = \max \frac{\log M}{n}$$

$\log M$ → no. of bits of messages sent

n ← no. of uses of the channel

$$= \max \frac{\log 2^{nI(X:Y)}}{n}$$

$$= \max_{p(x)} I(X:Y)$$

$p(x)$ ← over all possible input distributions