Information Theory & Coding

Nov 2nd, 2020

Last week : "Good news" for transmission of data :

- Given a channel $p(y|x)$,
  $R < C(P)$ then the rate R is "achievable".
- Proof by "random coding".

Slight variant of the proof:

fix p_X, (given $p_{Y|X}$ so, we have p_{XY})

choose x^n randomly as $X^n = (X_1 \dots X_n)$

iid $\sim p_X$

as codeword for $m=1$.

Choose other codewords

independent

$$M-1 \left\{ \begin{array}{l} \tilde{X}^n = (\tilde{X}_1 \dots \tilde{X}_n) \quad \text{iid} \sim p_X \\ \tilde{X}^n = (\qquad\qquad) \\ \text{etc.} \end{array} \right.$$

Send X^n over the channel, let the received sequence by Y^n

$Pr(X^n = x^n, Y^n = y^n) = P_X(x^n) P_{Y|X}(y^n|x^n)$

$Pr(\underbrace{\tilde{X}^n = \tilde{x}^n}, Y^n = y^n) = P_X(\tilde{x}^n) P_Y(y^n)$

the decoder computes a score of each message in the following way

$S_1 = \underline{P_{Y|X}(Y^n | X^n)} / P_Y(Y^n)$

$S_2 = P_{Y|X}(Y^n | \tilde{X}^n) / P_Y(Y^n)$

$\vdots$

$S_M = P_{Y|X}(Y^n | \tilde{\tilde{X}}^n) / P_Y(Y^n)$.

pick a threshold t and declare

$\hat{m} = 1$ if $S_1 \geq t$ & $S_2 < t, \ldots S_M < t$

$\hat{m} = 2$ if $S_2 \geq t$ & $S_1 < t, \ldots S_M < t$

$\hat{m}$ if $S_{\hat{m}} \geq t$ and $S_i < t$ $i \neq \hat{m}$

if no such $\hat{m}$, then output $\hat{M} \sim$ uniform$\{1 \ldots, M\}$.

$P(\text{Error}) \leq Pr(S_1 < t) + \underline{Pr(S_2 \geq t)} + \ldots + \underline{Pr(S_M \geq t)}$

$= Pr(S_1 < t) + (M-1) Pr(S_2 \geq t)$.

$$\frac{1}{n}\log S_1 = \frac{1}{n}\log \frac{p(Y^n | X^n)}{p(Y^n)}$$

with high prob

$$= \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(Y_i | X_i)}{p(Y_i)} \approx E \log \frac{p(Y_1 | X_1)}{p(Y_1)}$$

(X_i, Y_i) are iid $\sim p_{XY}$

$$E\left[\log \frac{p(Y|X)}{p(Y)}\right] = \sum_{x,y} p_{XY}(x,y) \log \frac{p_{Y|X}(y|x)}{p_Y(y)}$$

$$= I(X;Y)$$

Set $t = 2^{n(I(X;Y) - \varepsilon)}$, then

$$\Pr(S_1 < t) \to 0 \text{ as } n \to \infty.$$

L.L.N.

$$\boxed{\Pr(S_2 \geq t)}$$

$$= \Pr\left(\frac{1}{n}\log S_2 \geq \frac{1}{n}\log t\right)$$

$$= \Pr\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{p(Y_i | \tilde{X}_i)}{p(Y_i)} \geq I(X;Y) - \varepsilon\right).$$

$$\Pr(S_2 \geq t) \leq \frac{1}{t} E(S_2)$$

Markov Inequality

non-neg random variable $\quad S_2 = \frac{p(y^n \mid \tilde{X}^n)}{p(y^n)}$

$$E(S_2) = \sum_{\tilde{x}^n, y^n} p(\tilde{x}^n) \cancel{p(y^n)} \frac{p(y^n \mid \tilde{x}^n)}{\cancel{p(y^n)}}$$

$$= \sum_{\tilde{x}^n, y^n} \underbrace{p(y^n \mid \tilde{x}^n)} p(\tilde{x}^n) = 1$$

$$\Rightarrow \Pr(S_2 \geq t) \leq \frac{1}{t} = 2^{-n(I(X;Y) - \varepsilon)}$$

$$P(\text{Error}) \leq \Pr(S_1 < t) + \underbrace{(M-1)} 2^{-n[I(X;Y) - \varepsilon)}.$$

$$\leq \Pr(S_1 < t) + 2^{n(R - I(X;Y) + \varepsilon)}.$$

$$M = \lceil 2^{nR} \rceil$$

Now: with $R < C(P)$ we can find p_X s.t & $\varepsilon > 0$

$$I(X;Y) > R + \varepsilon$$

So $\quad R - I(X;Y) + \varepsilon < 0$

Now as we increase n we have

$$P(\text{Error}) \leq \underbrace{\Pr(S_1 \leq t)}_{\to 0 \text{ (LLN)}} + \underbrace{2^{n(\underbrace{\quad}_{<0})}}_{\to 0}$$

So we can make $P(\text{Error})$ as small as we wish by taking n large enough.

$\Rightarrow$ R is an achievable rate. //

Proof of the Markov inequality:

__Lemma__ if S is a ≥ 0 RV and $t \geq 0$

Then $\Pr(S \geq t) \leq \frac{1}{t} E[S]$.

__Pf__: $\Pr(S \geq t) = E[\mathbb{1}\{S \geq t\}]$

$$\mathbb{1}\{S \geq t\} \leq \frac{S}{t} \quad \text{so}$$

$$E(\quad) \leq E(\quad) \quad //$$

Consequences of Markov's inequality are many:

Ex. Z is a $\mathbb{R}$-valued RV

$e^{\lambda Z}$ is a ≥ 0 RV.

$$\Pr(Z \geq t) = \Pr(e^{\lambda Z} \geq e^{\lambda t}) \qquad \lambda > 0$$

$$\leq e^{-\lambda t} E[e^{\lambda Z}]$$

So $\Pr(Z \geq t) \leq \underbrace{\min_{\lambda > 0} e^{-\lambda t} E[e^{\lambda Z}]}$

Chernoff bound.

Channels with cost:

We are given a Memoryless channel $p(y|x)$ also a ~~cost~~ cost function on $\mathcal{X}$

$b: \mathcal{X} \to \mathbb{R}$. We wish to communicate

at ① high rate, ② low error probability ③ low cost.

Given $p(y|x)$ & $b(x)$ we say that

R is achievable at cost β if

$\forall \varepsilon > 0$, we can design $enc()$ & $dec()$

$enc: \{1..M\} \to \mathcal{X}^n$

s.t. ① $rate(enc) \geq R$

② $\hat{P}_e(enc, P, dec) < \varepsilon$

③ $\max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^{n} b(enc(m)_i) < \beta + \varepsilon$

Thm: given $p(y|x)$, $b: \mathcal{X} \to \mathbb{R}$ compute

$$C(P, \beta) = \max_{P_X :\ E[b(X)] \leq \beta} I(X;Y),$$

then any $R < C(P, \beta)$ is achievable at cost β.

Pf: Given $R < C(P, \beta)$ pick P_X, $\varepsilon > 0$

s.t. ①. $R + \varepsilon < I(X;Y)$

②. $E[b(X)] \leq \beta$

Then pick n large enough (we will see how large)

and $M = \lceil 2^{nR} \rceil$. and randomly choose enc(1), .., enc(M) as in the proof today.

Let us modify the decoder as follows:

Compute $S_1, .., S_M$ and declare

$\hat{m}$ if it is the only m s.t.

$S_{\hat{m}} \geq t$, $S_i < t$ $i \neq \hat{m}$,

$\frac{1}{n}\sum_{i=1}^{n} b(enc(\hat{m})_i) < \beta + \varepsilon$..

$$P(Error) \leq Pr(S_1 < t) + Pr(S_2 \geq t)(M-1)$$
$$+ Pr\left(\frac{1}{n}\sum_{i=1}^{n} b(enc)_i \geq \beta + \varepsilon\right).$$

the previous proof already shows that the 1st & 2nd terms $\searrow 0$ as $n \nearrow$. But the 3rd term:

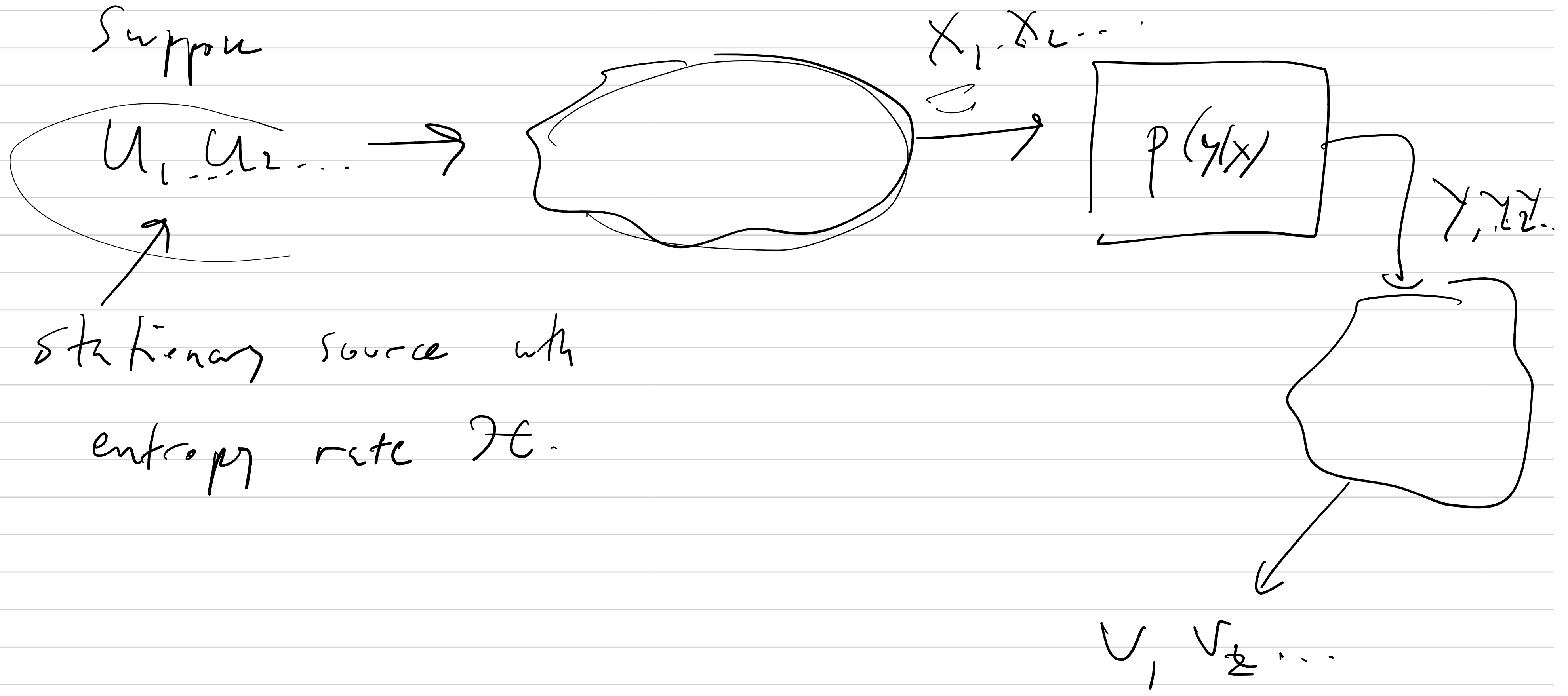
$$\underbrace{Pr\left(\frac{1}{n}\sum_{i=1}^{n} b(X_i) \geq \beta + \varepsilon\right)}_{LLN \rightarrow E[b(X_1)] \leq \beta} \searrow 0$$

by LLN.

//

Conversely; Thm:

Suppose



Let

s = # of source letter / channel use,

Suppose $\beta = \frac{1}{n}\sum_{i=1}^{n} E(b(X_i))$,

and suppose $p = \frac{1}{k}\sum_{i=1}^{k} \Pr(U_i \neq V_i)$.

Then:
$$\boxed{h_2(p) + p\log(|\mathcal{U}|-1)}$$
$$\geq \mathcal{H} - \frac{1}{s}C(P,\beta).$$

<u>Pf</u>: this will follow in exactly the same way as the "no cost" counterpart of this theorem we proved previously, once we show

$$\underline{\underline{I(X^n;Y^n) \le n\,C(P,\beta)}}.$$

to show this note that

$$I(X^n;Y^n) \le \sum_{i=1}^{n} \underbrace{I(X_i;Y_i)}_{f(p_{X_i},\, \underline{P})}$$

$$f(p,P) = I(X;Y)\Big|_{p_X = p,\ p_{Y|X} = P}.$$

Recall that $f(p_X, P)$ is concave in p_X..

Then $\underbrace{\frac{1}{n}\sum_{i=1}^{n}}\ f(p_{X_i}, \underline{P}) \le f(\underbrace{\frac{1}{n}\sum_{i=1}^{n} p_{X_i}}_{q_X}, \underline{P}).$

note that

$$\underbrace{\sum_{x} q_X(x)\, b(x)}_{E[b(X)]\big|_{X\sim q}} = \sum_{x}\frac{1}{n}\sum_{i=1}^{n} p_{X_i}(x)\, b(x)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\sum_{x} p_{X_i}(x)\, b(x)}_{E(b(X))\big|_{X\sim p_{X_i}}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(b(X_i)) = \beta$$

So $\underline{\underline{I(X^n; Y^n)}} \le \sum_{i=1}^{n} I(X_i; Y_i)$

$$= n \frac{1}{n} \sum_{i=1}^{n} f(P_{X_i}, P)$$

$$\le n\, f(q_X, P)$$

$$= n\, \underline{I(X;Y)}\Big|_{\substack{P_X = q_X \\ E(b(X)) = \beta}} \le n\, \underline{\underline{C(P, \beta)}}$$

//

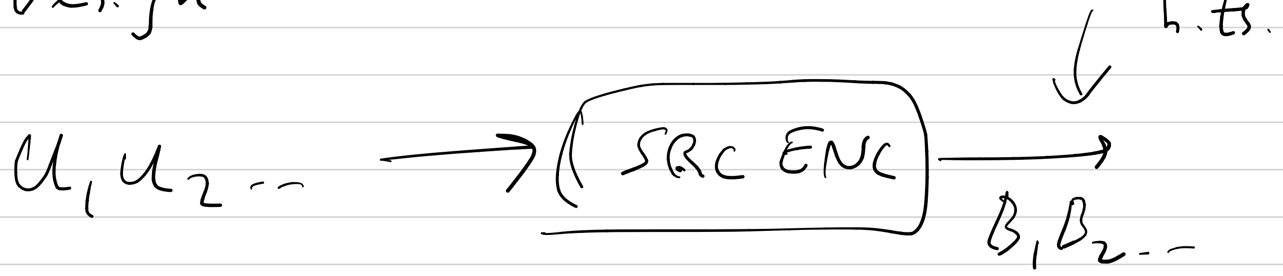
Corollary: $\underline{\mathcal{SH} > C(P,\beta)}$ is incompatible with reliable communication at cost β.

The "good news" theorems that show that reliable systems at all rates $< C$ exist suggests the following design principle:

- Given a source (stationary) $U_1, U_2 \dots$ producing ρ_s letter/seconds
- Given a channel $P(y|x)$ memoryless that we can use ρ_c times/second

— (ρ_s/ρ_c is the value of s in the prev. theorem).

- Design a source encoder

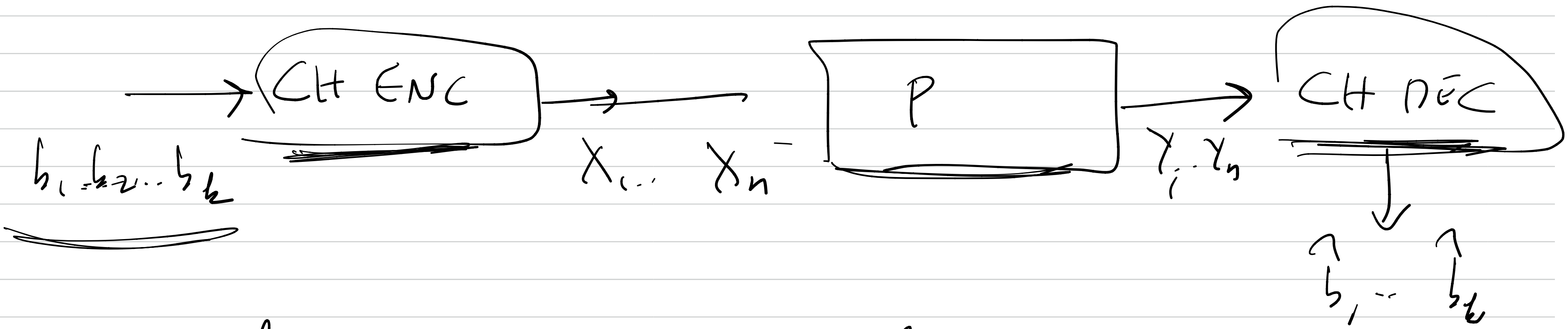


Such encoders exist for any $\varepsilon > 0$ and with

$\underline{\rho_s(\mathcal{H}+\varepsilon)}$ bits/sec produced at the output.

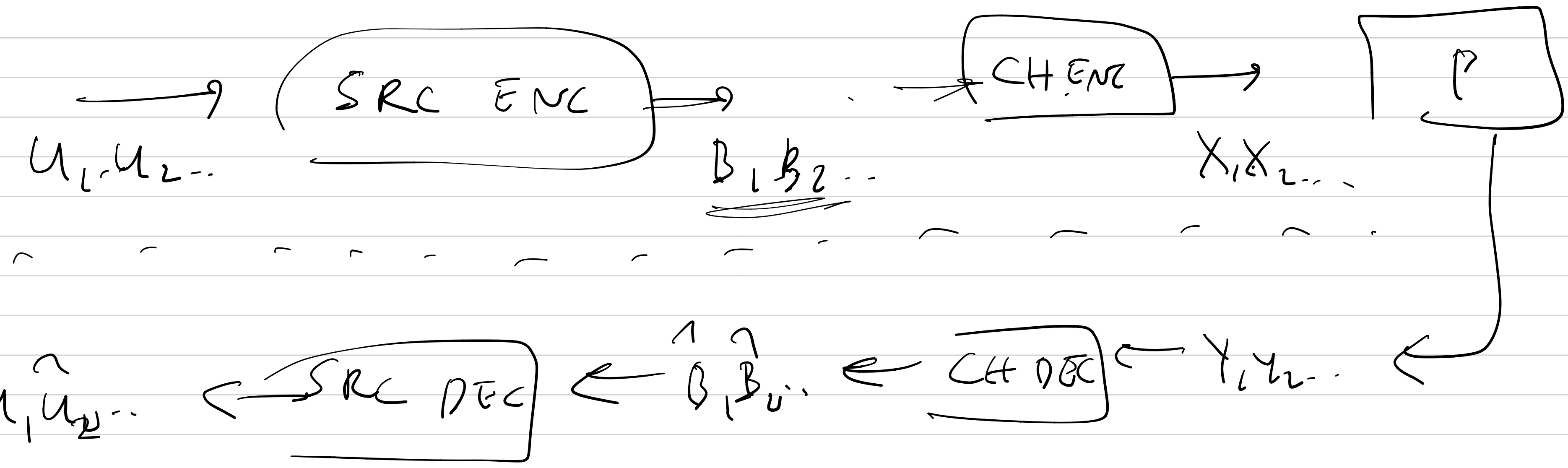
(also design the corresponding $\boxed{\text{SRC DEC}}$).

- For any $R < C$, we can design



with $\frac{k}{n} \geq R$ & $Pr(\hat{b}_1 \ldots \hat{b}_k \neq b_1 \ldots b_k) < \varepsilon$.

- now we can glue the designs.



as long as

$$\underbrace{\rho_s (H+\varepsilon) < \rho_c C}$$

the system will work with $P(\hat{U}_i \neq U_i) <$ small.

- The construction principle is an architectural one.
    - By source enc/dec turn any source into a bit stream.
    - By channel enc/dec turn any channel into a reliable bit pipe

It ~~might~~ could have been the case in a different universe that

$$U_1 U_2 \longrightarrow \text{Joint design} \longrightarrow X_1 X_2 X_2$$

allow us to reliably communicate at higher

# of source letters / channel use than systems

$U_1 U_2 \to$ SRC ENC $\to$ bits $\to$ CH ENC $\to X$'s

a "modular" design.