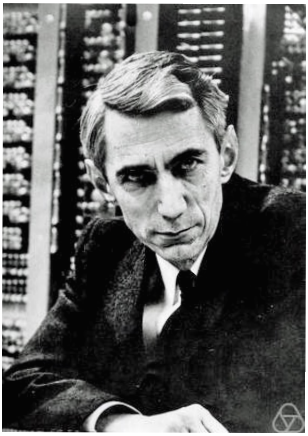
Channel coding theorem

Hangjun Cho.



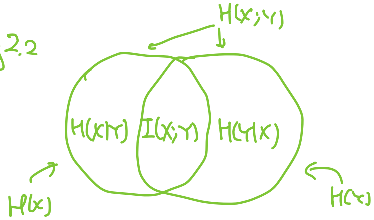
Claude Elwood Shannon
(April 30, 1916 – February 24, 2001)

$W \longrightarrow$ encoder $\xrightarrow{X^n}$ $p(y|x)$ Channel $\xrightarrow{Y^n}$ Decoder $\longrightarrow \hat{W}$

Def: A discrete channel is a system consisting of an input $\mathcal{X}$, output $\mathcal{Y}$, $p(y|x)$.

Information channel capacity: $C = \max_{p(x)} \underbrace{I(X;Y)}_{\text{mutual information.}}$

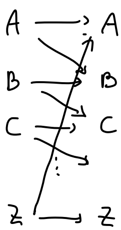
Fig2.2



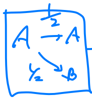
Thm 2.4.1.

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

ex) 7.1.3. Noisy Typewriter.



$$C = \max[H(Y) - H(Y|X)]$$



$H(Y|X) = \log_2 2 = 1$

$$= \max H(Y) - 1$$ Thm 2.6.4.

$$= \log 26 - 1$$

$$= \log 13$$

Thm 2.6.4. $H(X) \le \log|\mathcal{X}|$. equality holds iff X has a uniform dist. over $\mathcal{X}$.

§7.5 definition for theorem.

Def. • A discrete channel : $(\mathcal{X}, p(y|x), \mathcal{Y})$

finite set — prob. mass fcn.

• The nth extension of DMC : — discrete memoryless channel

$$(\mathcal{X}^n, p(y^n, x^n), \mathcal{Y}^n) \quad \text{where} \quad p(y_k | x^k, y^{k-1}) = p(y_k | x_k)$$

$y^n = (y_1, y_2, \cdots, y_n)$: vector notation.

• A channel without feedback if

$$p(x_k | x^{k-1}, y^{k-1}) = p(x_k | x^{k-1})$$

Def. An (M, n) code for $(\mathcal{X}, p(y|x), \mathcal{Y})$:

1. An index set : $\{1, 2, \cdots, M\} =: I^M$

2. encoding fcn : $X^n : I^M \to \mathcal{X}^n$ $\quad i \mapsto x^n(i)$ — codewords $\in$ codebook.

3. decoding fcn. : $g : \mathcal{Y}^n \to I^M$

Def. • Conditional prob. error. $\quad \lambda_i = P(g(Y^n) \neq i \mid X^n = x^n(i))$

• The maximal prob. of error. $\quad \lambda^{(n)} = \max_i \lambda_i$

• The average prob. of error. $\quad P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$.

- If the index W is chosen uniformly over I^M,

$$\boxed{P_e^{(n)} = P(W \neq g(Y^n))}$$

Def. The rate R of (M, n) Code : $R = \frac{\log M}{n}$

- A rate R is achievable if $\exists$ a seq. of $(\lceil 2^{nR} \rceil, n)$ Code, s.t. $\lambda^{(n)} \xrightarrow{n \to \infty} 0$

  $2^{nR} \in \mathbb{Z}$?

- The capacity of a DMC is the supremum of all achievable rates.

  (achievable capacity)

typical set

Def. The set $A_\varepsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$:

$$A_\varepsilon^{(n)} := \Big\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n :$$

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \varepsilon, \ \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \varepsilon, \ \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \varepsilon \Big\}$$

where $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$

$p(x_i, y_i) \neq p(x_i) p(y_i)$

Thm 7.6.1 (X^n, Y^n) : sequences drawn i.i.d. ad $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$

1. $\mathbb{P}((X^n, Y^n) \in A_\varepsilon^{(n)}) \to 1$ as $n \to \infty$

2. $|A_\varepsilon^{(n)}| \le 2^{n(H(X,Y)+\varepsilon)}$

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n) p(y^n)$ $\neq p(x^n, y^n)$, then $\mathbb{P}((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \le 2^{-n(I(X;Y) - 3\varepsilon)}$

   For $n \gg 1$, $\mathbb{P}((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \ge (1-\varepsilon) 2^{-n(I(X;Y)+3\varepsilon)}$

Thm. (Channel Coding Theorem.) For a DMC,

"$R < C \implies \exists (2^{nR}, n)$ code with $\lambda^{(n)} \to 0$ as $n \to \infty$"

Conversely, if $\exists$ a seq. of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, $R \le C$

pf. (Achievability) Fix $p(x)$ distribution. generate $(2^{nR}, n)$ code at random and

$$p(x^n) = \prod_{i=1}^{n} p(x_i).$$

which gives a code book.

$$C = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

the 1st message goes to n-vector.

Note that $\mathbb{P}(C) = \prod_{\omega=1}^{2^{nR}} p(x^n(\omega)) = \prod_{\omega=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(\omega))$

seq. of events:

1. C is generated and $p(x)$.

2. C is revealed to encoder and decoder
   $\downarrow$
   know $p(y|x)$

3. W is chosen and a uniform dist: $\mathbb{P}(W=\omega) = 2^{-nR}$, $\omega = 1, 2, \cdots, 2^{nR}$.

4. $X^n(\omega)$ is sent.

5. decoder receives a seq. Y^n and $P(y^n | x^n(\omega)) = \prod_{i=1}^{n} p(y_i | x_i(\omega))$.

6. The receiver decode the index:

$$\hat{\omega} = g(Y^n) \quad \text{if} \begin{cases} (X^n(\hat{\omega}), Y^n) \text{ is jointly typical.} \\ \nexists\, \omega' (\neq \hat{\omega}) \text{ s.t. } (X^n(\omega'), Y^n) \in A_\varepsilon^{(n)}. \end{cases}$$

error. when no such $\hat{\omega}$ exists or there is more than one such.

7. $\exists$ decoding error if $\hat{\omega} \neq \omega$. $\mathcal{E} = \{\hat{\omega} \neq \omega\}$.

(Study of prob. of error)

Calculate the ave. prob. of error, averaged over all codes.

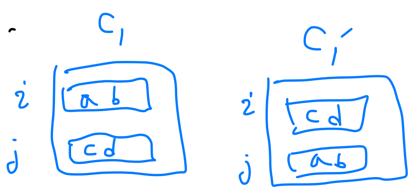
$$P(\mathcal{E}) = \sum_{C} P(C) P_e^{(n)}(C)$$

$P_e^{(n)} = P(W \neq g(Y^n))$

$$= \sum_{C} P(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C)$$

$$= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_{C} P(C) \lambda_\omega(C).$$

By symmetry of code construction, $\sum_{C} P(C) \lambda_\omega(C)$ does not depend on W.



WOLG ass. $W=1$ was sent.

$$\mathbb{P}(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{c} P(c)\lambda_1(c).$$

$$= \sum_{c} P(c)\lambda_1(c) = \mathbb{P}(\mathcal{E} \mid W=1).$$

Define $E_i = \{(X^n(i), Y^n) \in A_\varepsilon^{(n)}\}$ for $i = \{1, 2, \dots, 2^{nR}\}$

$\boxed{W=1}$

An error occurs:

- E_1^c $(X^n(1), Y^n)$ is not jointly typical.
- $E_2 \cup \cdots \cup E_{2^{nR}}$ Wrong codeword is jointly typical.

Y^n is the result of sending the first code word $X^n(1)$.

$$\mathbb{P}(\mathcal{E} \mid W=1) = \mathbb{P}(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}})$$

$$\leq P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i)$$

$P(A \cup B) \leq P(A) + P(B)$

$$\leq \underbrace{\varepsilon}_{1} + \underbrace{\sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y) - 3\varepsilon)}}_{3} \quad \text{by joint AEP.}$$

3 $X^n(1)$ and $X^n(i)$: indep. for $i \neq 1$.

$$= \varepsilon + (2^{nR} - 1) \cdot 2^{-n(I(X;Y) - 3\varepsilon)}$$

$$\leq \varepsilon + 2^{-n(I(X;Y) - 3\varepsilon - R)} \; (*) \leq 2\varepsilon.$$

(if $R < I(X;Y) - 3\varepsilon$ and $n \gg 1$.)

$\therefore$ For $R < I(X;Y)$, $\forall \varepsilon, \exists n$ s.t. $\mathbb{P}(\mathcal{E}) \leq 2\varepsilon$.

(Finding a Codebook we want)

1. Choose $p(x)$ that maximize $I(X;Y)$, and call it $p^*(x)$.

$$R < I(X;Y) \Rightarrow R < C = \max_{p(x)} I(X;Y)$$

2. $\exists C^*$ s.t. $P_e^{(n)}(C^*) \le 2\varepsilon$ $\leftarrow$ $P(\varepsilon) = \sum_c P(c) P_e^{(n)}$

3. $\boxed{\text{Aim : } \lambda^{(n)}(C^*) \longrightarrow 0}$

$\Lambda^* = \lambda_{\frac{n}{2}}$

$$2\varepsilon \ge \underbrace{\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*)}_{\Lambda^*} = \frac{1}{2^{nR}} \left(\underbrace{\sum_{\Lambda^* > \lambda_i} \lambda_i}_{\text{best half}} + \underbrace{\sum_{\Lambda^* < \lambda_i} \lambda_i}_{\text{worst half}} \right)$$

For best half, $\lambda_i \le 4\varepsilon$ for all i in best half.

If not $\Lambda^* \ge 4\varepsilon$, $\frac{1}{2^{nR}} \sum_{\Lambda^* < \lambda_i} \lambda_i > \frac{1}{2} \cdot 4\varepsilon = 2\varepsilon$ ※.

Throw away the worst half of codeword in C^*.

And take 2^{nR-1} codewords. (best half of C^*)

New Rate $\frac{1}{n} \cdot \log(2^{nR-1}) = R - \frac{1}{n}$

and $\lambda^{(n)} = \max \lambda_i \le 4\varepsilon$.

///

(Converse) "if $\exists$ a seq. of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, $R \le C$"

(special case)

Supp. $P_e^{(n)} = 0$, Ass: W: uniformly dist. over $\{1, 2, \cdots, 2^{nR}\}$.

$$
\begin{aligned}
nR &\stackrel{\sim}{=} H(W) \\
&= \cancel{H(W|Y^n)} + I(W; Y^n) \qquad 0 \ (\because P_e^{(n)} = 0;\ \hat{W} = g(Y^n) \to W) \\
&\le I(X^n; Y^n) \qquad : \text{Data processing}.\ W \to X^n \to Y^n. \\
&\le \sum_{i=1}^{n} I(X_i; Y_i) \qquad 7.9.2 \\
&\le nC \qquad \text{def. of } C
\end{aligned}
$$

Fig 2.2

$\therefore R \le C$.

7.9.1 Lemma (Fano's inequality)

For a DMC, W: uniformly distributed. $H(X^n|Y^n) \le 1 + P_e^{(n)} \cdot n \cdot R$

7.9.2. For DMC, $I(X^n; Y^n) \le nC$.

pf.
$$
\begin{aligned}
I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum H(Y_i | Y_1, \cdots, Y_{i-1}, X^n) \\
&= H(Y^n) - \sum H(Y_i | X_i) \qquad \leftarrow \text{DMC}. \\
&\le \sum H(Y_i) - \sum H(Y_i | X_i) \\
&= \sum I(X_i; Y_i) \le nC.
\end{aligned}
$$

- Converse of the Coding theorem.

"if $\exists$ a seq. of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, $R \le C$"

pf. Since $P_e^{(n)} \le \lambda^{(n)}$, $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$.

Ass. W : uniform dist with $P_e^{(n)} = P(\hat{W} \neq W)$.

$$nR = H(W) = H(W|Y^n) + I(W;Y^n)$$

$$\le H(W|Y^n) + I(X^n(W);Y^n) \qquad W \to X^n \to Y^n$$

$$\le 1 + P_n^{(e)} \cdot nR + I(X^n;Y^n)$$

$$\le 1 + P_n^{(e)} nR + nC$$

Dividing by n, $R \le \frac{1}{n} + P_n^{(e)} R + C$.

As $n \to \infty$, $R \le C$. ///

· Equality in the Converse

We consider the special case $P_n^{(e)} = 0$: zero error code

$\boxed{W \to \text{code}}$

$$nR \overset{\checkmark}{=} H(W)$$

$$\overset{\checkmark}{=} H(\cancel{W|\hat{W}}) + I(W;\hat{W})$$

0 ($\because P_e^{(n)} = 0$)

Fig 2.2. X^n must be distinct.

$$\overset{\checkmark}{\le} I(X^n(W); Y^n)$$

data processing ineq. : $W \to X^n \to Y^n \to \hat{W}$.

$I(Y^n; X^n(W) | W) = 0 = I(X^n; Y^n | \hat{W})$

$$= H(Y^n) - H(Y^n | X^n)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | X_i) \quad \leftarrow \text{DMC}$$

$$\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) \quad \leftarrow Y_i : \text{indep.}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \quad \leftarrow \text{by def.}$$

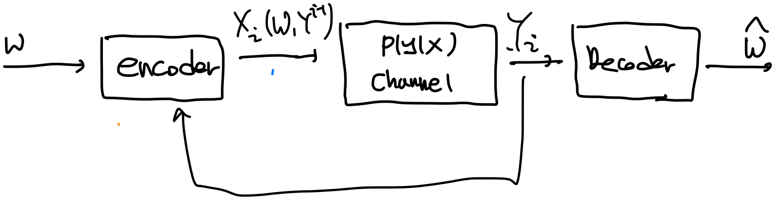
$$\le nC \quad \leftarrow \text{by def}$$

$X_i \sim p^*(x)$.

$||$

$\arg\max I(X;Y)$

§7.12 Feedback Capacity



- $(2^{nR}, n)$ feedback code : a sequence of mappings $x_i(W, Y^{i-1})$,

  seq of decoding fcns $g: \mathcal{Y}^n \to \{1, 2, \cdots, 2^{nR}\}$

The capacity with feedback, C_{FB}, of a DMC

: a supremum of all rates achievable by feedback codes.

7.12.1. $C_{FB} = C = \max_{p(x)} I(X; Y)$.

pf. A non feedback code is a special code of a feedback code

$$C_{FB} \geq C.$$

Aim: $C_{FB} \leq C$.

$$nR = H(W) = H(W|\hat{W}) + I(W; \hat{W})$$

(*) $$\leq 1 + P_e^{(n)} nR + I(W; \hat{W})$$

$$\leq 1 + P_e^{(n)} nR + I(W; Y^n) \quad \Leftarrow W \to Y^n \to \hat{W}.$$

$$I(W;Y^n) = H(Y^n) - H(Y^n|W)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y_1, Y_2, \dots, Y_{i-1}, W)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | \boxed{Y_1, Y_2, \dots, Y_{i-1}, W}, X_i) \qquad X_i = X_i(W, Y^{i-1})$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i | X_i) \qquad \text{DMC}$$

$$\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) \qquad \text{Chain rule} + 2.6.5 \quad H(X|Y) \le H(X)$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \qquad \text{def}$$

$$\le nC. \qquad \text{def}$$

from $(*)$, $nR \le P_e^{(n)} nR + 1 + nC$.

Dividing n, $n \to \infty$, $R \le C$.

$$C_{FB} \le C.$$

//

§7.13: Source channel separation theorem.

Two stage method.

$$\left(\begin{array}{lll} \text{Ch 5} & 5.4.2. \to & R_s > H \\ & & \quad \shortparallel \\ \text{ch 7.} & 7.7.1 \to & R_c < C. \end{array}\right) \to H(U) < C.$$

5.4.2 $\quad \dfrac{H(X_1,\cdots,X_n)}{n} \le L_n^* = \boxed{\dfrac{nL_n^*}{n}} \le \dfrac{H(X_1,\cdots,X_n)}{n} + \dfrac{1}{n}$

$H(\mathcal{X})$ $\qquad$ $nL_n^* = k$ $\qquad$ R $\qquad$ $H < R$

$k = \log M \to M = 2^k$, $k = nL_n^*$. $H(U) < C$.

Source-channel separation thm:

two-stage process is as good as any other method.

Source V. Process produced by V satisfies AEP. ex) i.i.d. ← Stationary.

Def. Prob. of error : $\mathbb{P}(V \ne \hat{V}) = \sum_{y^n}\sum_{x^n} p(v^n)\, p(y^n|x^n(v^n))\, I(g(y^n) \ne v^n)$

$V^n \to$ [en.] $\xrightarrow{X^n(V^n)}$ [ch $p(y|x)$] $\xrightarrow{Y^n}$ [de.] $\to \hat{V}^n$

Thm (Joint Source-Channel Coding thm) : DMC.

If $V_1, V_2, \cdots, V_n$ is a finite alphabet stochastic process that satisfies the AEP and $H(V) < C$,

$\exists$ source-channel code with $\Pr(\hat{V} \ne V) \to 0$

Conversely, for any stationary stochastic process, if $H(V) > C$, the probability of error is bounded away from zero.

Aim: $R < C \Rightarrow P_e^{(n)} \to 0$
$\quad$ ↳ $H(V)$. replace.

pf. Achievability ($\longrightarrow$)

AEP assumption on the process : $|A_\varepsilon^{(n)}| < 2^{n(H(V)+\varepsilon)}$ $\qquad$ (thm 3.1.2 − 3)

We index all seq. in $A_\varepsilon^{(n)}$ with $n(H(V)+\varepsilon)$ bits

compress.

3.1.2-2 : $\mathbb{P}(A_\varepsilon^{(n)}) > 1-\varepsilon$.

From ch. coding thm, We can transmit the indices with the arbitrary small prob of error if $H(U)+\varepsilon = R < C$

Can reconstruct V^n to agree with the transmitted sequence with high prob.

(why. Input size = n = output size)

Precisely,

$$P(V^n \neq \hat{V}^n) \le \mathbb{P}(V^n \notin A_\varepsilon^{(n)}) + \mathbb{P}(g(Y^n) \neq V^n \mid V^n \in A_\varepsilon^{(n)})$$
$$\le \varepsilon + \varepsilon = 2\varepsilon$$

$\therefore H(U) < C$.

Converse: "$\mathbb{P}(V^n \neq \hat{V}^n) \to 0$" $\Rightarrow H(U) \le C$

Given the following encoder and decoder $X^n(V^n): \mathcal{U}^n \to \mathcal{X}^n$, $g_n(Y^n): \mathcal{Y}^n \to \mathcal{U}^n$

By Fano's ineq.

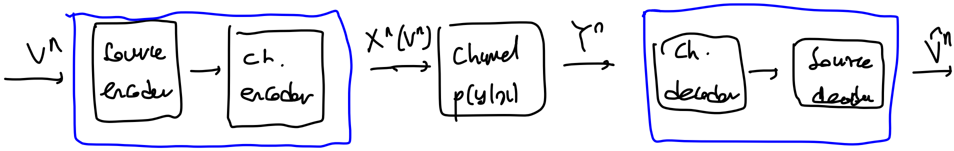
$$H(V^n \mid \hat{V}^n) \le 1 + P_r(\hat{V}^n \neq V^n) \cdot \log \underbrace{|\mathcal{U}^n|}_{= |\mathcal{U}|^n \text{ (Only seq.)}} = 1 + P_r(\hat{V}^n \neq V^n) \, n \cdot \log |\mathcal{U}|$$

$$H(U) \le \frac{H(X_1, \cdots, X_n)}{n}$$

Ch 4. $H(U) = \lim \frac{H(U_1, \cdots, U_n)}{n}$, #4.6 + dec. seq.

$$= \frac{H(V^n)}{n}$$
$$= \frac{1}{n} H(V^n \mid \hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n)$$
$$\le \frac{1}{n}(1 + \mathbb{P}(\hat{V}^n \neq V^n) \, n \log |\mathcal{U}|) + \frac{1}{n} I(V^n; \hat{V}^n) \quad \text{Fano}$$
$$\le \frac{1}{n}(1 + \mathbb{P}(\hat{V}^n \neq V^n) \, n \log |\mathcal{U}|) + \frac{1}{n} I(X^n; Y^n) \quad \text{Data processing}$$
$$\le \frac{1}{n}(1 + \mathbb{P}(\hat{V}^n \neq V^n) \, n \log |\mathcal{U}|) + C \quad \text{Memoryless}$$

As $n \to \infty$, $\therefore H(U) \le C$ ///



Summary. Data compression thm: a consequence of the AEP.

Data transmission thm: a consequence of the joint AEP.