AEP

1) By WLLN

$$-\frac{1}{n}\log P(X^n) \to -\mathbb{E}[\log P(X)] = H(X) \quad \text{in prob.}$$

give $\varepsilon > 0$, $\exists\, n_1$, for $n > n_1$

$$P\Big(\underbrace{\left|-\frac{1}{n}\log(P(X^n) - H(X)\right| \geq \varepsilon}_{A_1}\Big) < \frac{\varepsilon}{3}$$

Similarly, WLLN

$$-\frac{1}{n}\log P(X^n) \to -E[\log P(Y)] = H(Y) \quad \text{in prob}$$

$$-\frac{1}{n}\log P(X^n, Y^n) \to -E[\log P(X,Y)] = H(X,Y) \quad \text{in prob.}$$

$\exists\; n_2, n3$

for all $n \geq n_2$

$$P\Big(\underbrace{\left|-\frac{1}{n}\log P(Y^n) - H(Y)\right| \geq \varepsilon}_{A_2}\Big) < \frac{\varepsilon}{3}$$

for all $n \geq n_3$

$$P\Big(\underbrace{\left|-\frac{1}{n}\log P(X^n, Y^n) - H(X,Y)\right| \geq \varepsilon}_{A_3}\Big) < \frac{\varepsilon}{3}$$

choose $n > \max\{n_1, n_2, n_3\}$

$$P(A_1 \cup A_2 \cup A_3) \leq \sum_{i=1}^{3} P(A_i) = \varepsilon$$

"union bound"

$$A_\varepsilon^{(n)} = A_1^c \cap A_2^c \cap A_3^c = (A_1 \cup A_2 \cup A_3)^c$$

$$P(\{A_\varepsilon^{(n)}\}^c) \cancel{= 1 - P(A_1 \cup A_2 \cup A_3)}$$

$$\leq \varepsilon$$

$P(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$. for n suff. large.

② $$1 = \sum p(x^n, y^n)$$
$$\geq \sum_{A_\varepsilon^{(n)}} p(x^n, y^n)$$
$$\geq |A_\varepsilon^{(n)}| \, 2^{-n(H(X,Y)+\varepsilon)}$$
$$|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$$

③ $\tilde{x}^n, \tilde{y}^n$ are independent, having the same marginal as x^n, y^n, then

$$P((\tilde{x}^n, \tilde{y}^n) \in A_\varepsilon^{(n)}) = \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n)$$
$$\leq 2^{n(H(X,Y)+\varepsilon)} \cdot 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)}$$
$$= 2^{-n(\underbrace{-H(X,Y)+H(X)+H(Y)}_{I(X;Y)} - 3\varepsilon)}$$
$$= 2^{-n(I(X;Y)-3\varepsilon)}$$

For sufficient large n, $P(A_\varepsilon^{(n)}) \geq 1-\varepsilon$

$$1-\varepsilon \leq \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n, y^n)$$
$$\leq |A_\varepsilon^{(n)}| \, 2^{-n(H(X,Y)-\varepsilon)}$$

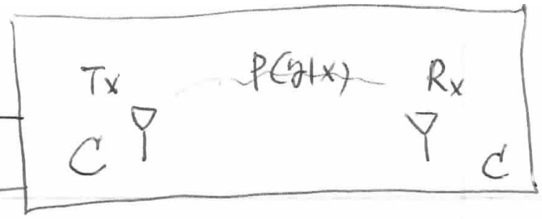
and $$|A_\varepsilon^{(n)}| \geq (1-\varepsilon) 2^{n(H(X,Y)-\varepsilon)}$$

$$P((\tilde{x}_n, \tilde{y}_n) \in A_\varepsilon^{(n)}) = \sum_{A_\varepsilon^{(n)}} p(x^n)p(y^n)$$
$$\geq (1-\varepsilon) 2^{n(H(X,Y)-\varepsilon)} 2^{-n(H(X)+\varepsilon)} 2^{-n(H(Y)+\varepsilon)}$$
$$= (1-\varepsilon) 2^{-n(I(X;Y)+3\varepsilon)}$$

① $R<C$, exi...

② if $P_e^{(n)} \to 0$, must $R<C$

Channel Coding theorem



(1) prove that $R<C$ are achievable

Fix $p(x)$. Generate $(\lceil 2^{nR} \rceil, n)$ code at random according to $P(X)$.

Generate 2^{nR} codewords independently according to

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

<u>Codebook</u>: Consists of 2^{nR} codewords

$$C = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ \vdots & \vdots & & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{bmatrix} \leftarrow \text{codeword}$$

each entry iid $\sim P(x)$

$$Pr(C) = \prod_{w=1}^{\overset{M}{2^{nR}}} \underbrace{\prod_{i=1}^{n}}_{\text{each symbol}} P(X_i(w))$$

<u>Codebook</u> known to both sides

<u>Channel</u> known to both sides (distribution)

<u>Message</u> uniform distribution

$$P(W=w) = 2^{-nR}$$

$$w = 1, 2, \cdots, 2^{nR}$$

<u>Rx receives</u>

$$P(y^n | x^n(w)) = \prod_{i=1}^{n} P(y_i | x_i(w))$$

at Rx

Joint typical decoding

easy to analyze, asymptotically optimal

(ML decoding optimal, but not easy to analyze)

decoder find $\hat{w}$ ① if $(x^n(\hat{w}), y^n)$ is jointly typical.

② "no confusion"

no other index $w' \neq \hat{w}$

s.t. $(x^n(w'), y^n) \in A_\epsilon^{(n)}$

error when (a) cannot find

(b) find more than one

Decoding error: $\mathcal{E} = \{\hat{w} \neq w\}$

Analysis

find prob. of error (not for a single code),

but over all codes generated at random.

Proof Let w be drawn uniformly from $\{1, 2, \cdots, 2^{nR}\}$

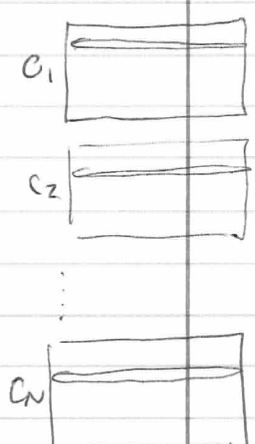
use joint typical decoding to find $\hat{w}(y^n)$

Let $\mathcal{E} = \{\hat{w}(y^n) \neq w\}$ denote error event

Prob. averaged over all codewords in the codebook, and over all codebook

$\lambda_i = P\{g(y^n) \neq i \mid X^n = X^n(i)\}$

Intuition: two things are random:
$C \sim p(x)$
$y^n \mid x_n \sim p(y|x)$



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

$$= \sum_{C} Pr(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

exchange order

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \underbrace{\sum_{C} Pr(C)\lambda_w(C)}_{\sum_{C} Pr(C)\lambda_1(C)}$$

By symmetry, avei prob. of err does not depend on particular index sent.

$$= \sum_{C} Pr(C)\lambda_1(C)$$

$$= P(\mathcal{E} \mid \underbrace{w=1}_{\text{assume message 1 was sent}})$$

Define joint typical event.

$$E_i = \{ (X^n(i), Y^n) \text{ is in } A_\varepsilon^{(n)} \}$$
$$i = 1, \cdots 2^{nR}$$

Now fix Y^n to be the outcome when $X^n(1)$ was sent.

$$\Rightarrow P(\mathcal{E} \mid W=1) = P(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} \mid W=1)$$

$$\leq \underbrace{P(E_1^c \mid W=1)}_{① \to 0, \text{ by joint AEP}} + \sum_{i=2}^{2^{nR}} \underbrace{P(E_i \mid W=1)}_{② \leq 2^{-n(I(X;Y)-3\varepsilon)} \text{ joint AEP}}$$

① $P(E_1^c \mid W=1) \leq \varepsilon$, for n sufficiently large.

② $X^n(1)$ and $X^n(i)$ indpt. for $i \neq 1$

$\Rightarrow X^n(i)$ and Y^n are indept.

joint AEP

$\Rightarrow$

$$P(E_i \mid W=1) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

Finally: $P_r(\mathcal{E}) \leq \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\varepsilon)}$

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

$$\leq \varepsilon + \cancel{2^{3n\varepsilon}} 2^{-n(\underbrace{I(X;Y) - R}_{\text{if } \geq 0} \cancel{-3\varepsilon})} \leq 2\varepsilon$$

if for n sufficiently large and $R < I(X;Y) - 3\varepsilon$.

To strengthen the result,

1. choose $P(x)$ to be $p^*(x)$

$$p^*(x) = \arg\max_{P(x)} I(X;Y)$$

$\Rightarrow$ $R < I(X;Y)$ becomes

$R < C$

2. get ride of average over codebook.

Since the ave. over codebook is $\leq 2\varepsilon$

exists at least one codebook C^* w.

small prob of err.

$P(\varepsilon \mid C^*) \leq 2\varepsilon$

C^* can be found by (at least exaustive search)

3.

~~Setup~~

- Throw away the worst half of the codewords in the best codebook C^*.

Since arithmetic average prob. of error $P_e^{(n)}(C^*)$ for this code is less than 2ε

$$P(\mathcal{E}|C^*) \le \frac{1}{2^{nR}} \sum \lambda_i(C^*) \le 2\varepsilon$$

$\Rightarrow$ at least half the indices i and their $X^n(i)$ have $\lambda_i \le 4\varepsilon$

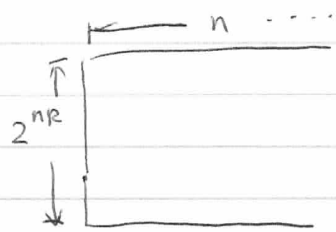
$\Rightarrow$ ~~less~~ the best half of the codewords have max prob err $\lambda^{(n)} \le 4\varepsilon$

if we reindex these codewords, we have 2^{nR-1} codewords, rate from R to $R-\frac{1}{n}$

which is negligible for large n.

Special case (help with proof of converse)

- $P_e^{(n)} = 0$ implies $R \le C$
- $(2^{nR}, n)$ code



with zero Pe

$$\Rightarrow H(W \mid Y^n) = 0$$

- assume W is uniformly distributed

$$nR = H(W) = \underbrace{H(W \mid Y^n)}_{=0} + I(W; Y^n)$$

$$= I(W; Y^n)$$

$$\le I(X^n; Y^n)$$

$$\le \underbrace{\sum_{i=1}^{n} I(X_i, Y_i)}_{\text{due to discrete memoryless assumption}}$$

(data processing inequality)

$$\le nC \quad \left(C = \max_{p(x)} I(X, Y) \ge I(X_i, Y_i) \right)$$

hence for zero-Pe $(2^{nR}, n)$ code,

$$R \le C.$$

· We can prove (*): for DMC

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

<u>Proof</u>

$$\begin{aligned}
I(X^n; Y^n) &= H(Y^n) - H(Y^n \mid X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid Y_1, \cdots, Y_{i-1}, X^n) \\
&\qquad \text{(chain rule)} \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid X_i) \\
&\qquad \text{(DMC, no feedback)} \\
&\leq \underbrace{\sum_{i=1}^{n} H(Y_i)}_{\text{union bound}} - \sum_{i=1}^{n} H(Y_i \mid X_i) \\
&= \sum_{i=1}^{n} I(X_i; Y_i)
\end{aligned}$$

Proof of converse

- new ingredient: Fano's inequality

$$P_e \geq \frac{H(X|Y) - 1}{\log |X|}$$

- Let's setup the problem

the index W uniformly distributed on

$$\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$$

$$W \xrightarrow[f]{} X^n(W) \xrightarrow[p(y|x)]{} Y^n \xrightarrow[g]{} \hat{W}$$

- Define probability of error

$$P(\hat{W} \neq W) = \frac{1}{2^{nR}} \sum_i \lambda_i = P_e^{(n)}$$

probability of err for ith codeword, fixed codebook

$$\lambda_i = P(g(Y^n) \neq i \mid X^n = x^n(i))$$

$$= \sum_{y^n} p(y^n | x^n(i)) \, I(g(y^n) \neq i)$$

- Fano's inequality says that

$$P_e^{(n)} \geq \frac{H(W|\hat{W}) - 1}{\underbrace{\log M}_{= nR}}$$

$$\Rightarrow \quad H(W|\hat{W}) \leq 1 + P_e^{(n)} nR$$

<u>Goal</u> show that any sequence of $(2^{nR}, n)$ code with $\lambda^{(n)} \to 0$, must have $R < C$.

$$\lambda^{(n)} = \max_{i \in \{1,2,\cdots,M\}} \lambda_i$$

- Let W be uniformly distributed over $\{1, 2, \cdots, 2^{nR}\}$

- $P(\hat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \Sigma_i \lambda_i$

  $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$, as $n \to \infty$

- $$\begin{aligned} nR &= H(W) \\ &= H(W|\hat{W}) + I(W;\hat{W}) \\ &\le \underbrace{1 + P_e^{(n)} nR}_{\text{Fano's inequality}} + I(W;\hat{W}) \\ &\le 1 + P_e^{(n)} nR + I(X^n;Y^n) \\ &\quad \text{(data processing inequality)} \\ &\le 1 + P_e^{(n)} nR + nC \\ &\quad \text{(channel capacity)} \end{aligned}$$

- Divide both sides by n.

$$R \le \frac{1}{n} + P_e^{(n)} R + C$$

letting $n \to \infty$, $P_e^{(n)} \to 0$

$$\boxed{R \le C}$$

On the other hand, we can write

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

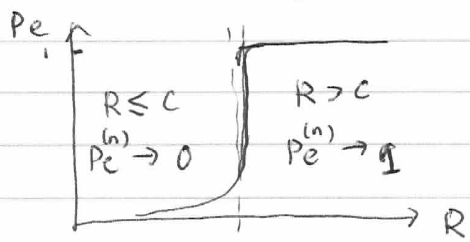
if $R > C$, $\frac{C}{R} < 1$

for large n, $P_e^{(n)}$ is bounded away from 0.

Hence if $R > C$, we cannot achieve an arbitrarily low probability of error.

$\blacksquare$

- This is the weak converse
- Strong converse: $P_e^{(n)} \to 1$ exponentially if $R > C$.



- Equality in the converse to the channel coding theorem

→ How to find capacity achieving codes?

$$nR = H(W)$$
$$= \underbrace{H(W|\hat{W})}_{=0 \text{ zero } Pe} + I(W;\hat{W})$$
$$= I(W;\hat{W})$$
$$\underset{(a)}{\leq} I(X^n(W);Y^n) \quad \text{(data processing}$$

Data processing inequality

$$X \to Y \to Z, \Rightarrow I(X;Y) \geq I(X;Z)$$

equality iff

$I(X;Y|Z) = 0$ (i.e. $X \to Z \to Y$ also forms Markov chain)

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

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

$$\overset{(b)}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)$$

(~~union bound~~)

$$= \sum_{i=1}^{n} I(X_i; Y_i) \underset{(c)}{\leq} nC$$

(a) : equality iff $I(X^n; Y^n | W) = 0$

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

true when all code words are distinct & $\hat{w}$ is sufficient stats for decoding

(b) Y_i independen

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

∎

Capacity achieving zero-err code must have

1. distinctive codewords
2. distribution of Y_i must iid w.

$$p^*(y) = \sum_x p^*(x) p(y|x)$$

~~eg~~ capacity achieving example: noisy typewriter.