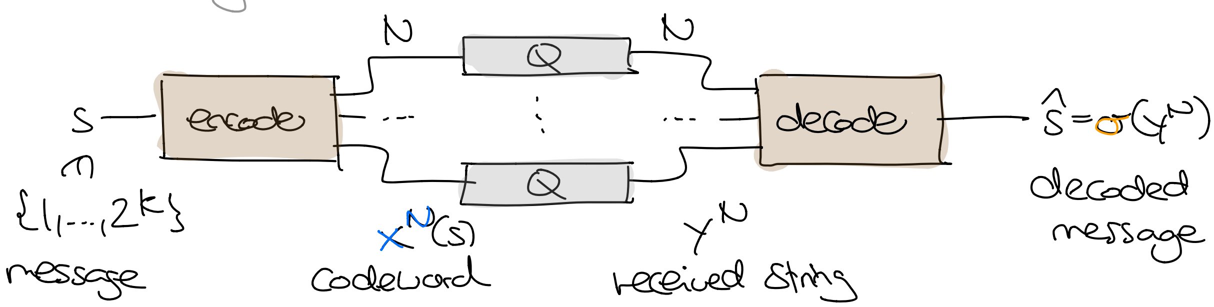
Proof of the Noisy Coding Theorem (§10)

Recall from Tuesday:



(N,K)-block code: $x^N: \{1,2,\dots,2^K\} \longrightarrow \mathcal{A}_X^N$

Decoder: $\sigma: \mathcal{A}_Y^N \longrightarrow \{\perp, 1, \dots, 2^K\}$

Figures of merit:

* rate: $R := \frac{K}{N}$ bits per channel use
* average prob. of (block) error for uniform $S \in \{1,\dots,2^K\}$:

$$p_B = \Pr(\hat{S} \neq S) = \frac{1}{2^K} \sum_{s=1}^{2^K} \sum_{\hat{s} \neq s} P(\hat{s}|s)$$

similarly for general $P(s)$

* maximal probability of (block) error:

$$p_{BM} = \max_s \Pr(\hat{S} \neq S \mid S = s) = \max_s \sum_{\hat{s} \neq s} P(\hat{s}|s) \geq p_B$$

$\boxed{\text{Shannon's noisy coding theorem:}}$ Let $Q(y|x)$ channel.

(A) Achievability:

If $R < C(Q)$: $\forall \delta > 0$: $\exists N_0 \; \forall N \geq N_0$: $\exists$ code with $\frac{K}{N} \geq R$ & $p_{BM} \leq \delta$

(B) Converse:

If $R > C(Q)$: $\exists \delta > 0$ $\exists N_0 \; \forall N \geq N_0$: $\nexists$ code with $\frac{K}{N} \geq R$ & $p_B \leq \delta$

"weak converse" (also true $\forall \delta$ but will not prove this

Proof of Achievability (A)

Main tool: Jointly typical set for $P(x,y)$:

$$J_{N,\varepsilon}(P) = \left\{ \begin{array}{c} (x^N, y^N) \text{ s.th. } x^N \in T_{N,\varepsilon}(P_x),\ y^N \in T_{N,\varepsilon}(P_y) \\ \text{and } (x^N, y^N) \in T_{N,\varepsilon}(P_{xy}) \end{array} \right\}$$

Properties:

⓪ For all $(x^N, y^N) \in J_{N,\varepsilon}$: $2^{-N(H(X)+\varepsilon)} \leq P(x^N) \leq 2^{-N(H(X)-\varepsilon)}$ etc

① $\# J_{N,\varepsilon} \leq 2^{N(H(XY)+\varepsilon)}$

② If $(X^N, Y^N) \overset{\text{IID}}{\sim} P(x,y)$: ← $(X_i, Y_i) \sim P$

$$\Pr\left((X^N, Y^N) \in J_{N,\varepsilon}\right) \longrightarrow 1 \text{ as } N \to \infty$$

③ If $\tilde{X}^N \overset{\text{IID}}{\sim} P(x)$ & $\tilde{Y}^N \overset{\text{IID}}{\sim} P(y)$ independent: ← $\tilde{X}_i, \tilde{Y}_i$ independent

$$\Pr\left((\tilde{X}^N, \tilde{Y}^N) \in J_{N,\varepsilon}\right) \leq 2^{-N(I(X:Y)-3\varepsilon)}$$

Pf: $$\text{LHS} \overset{\text{independence}}{=} \sum_{(x^N, y^N) \in J_{N,\varepsilon}} P(x^N) P(y^N) \overset{⓪+①}{\boxed{\leq}} \# J_{N,\varepsilon} \cdot 2^{-N(H(X)-\varepsilon)}\, 2^{-N(H(Y)-\varepsilon)}$$

$$\leq 2^{-N(I(X:Y)-3\varepsilon)}$$ □

Enough to prove: For all $P(x)$, $R < \underbrace{I(X:Y)}_{\text{equal to capacity for suitable } P(x)}$, $\delta > 0$: $\exists$ sequence of (N,K)-block codes (one for each N) with $\frac{K}{N} \geq R$ s.th. $P_B \xrightarrow{N\to\infty} 0$

$K = K(N)$

Can always upgrade to $P_{B,m}$ via expurgation w/o changing rate much (→ last time)

Key idea: Choose code at random!

Random code: Let $K = \lceil NR \rceil$ and choose 2^K codewords at random:

$$\begin{matrix} X^N(1) = & X_1(1) & X_2(1) & \cdots & X_N(1) \\ \vdots & \vdots & & & \vdots \\ X^N(2^K) = & X_1(2^K) & X_2(2^K) & \cdots & X_N(2^K) \end{matrix} \quad \overset{\text{IID}}{\sim} P(x)$$

Codeword by codeword, letter by letter

$\hookrightarrow$ (N,K)-code with $\frac{K}{N} \geq R$

Typical set decoder: (deterministic)

$$\sigma(y^N) = \begin{cases} \hat{S} & \text{if exactly one } \hat{S} \text{ s.th. } (X^N(\hat{S}), Y^N) \in J_{N,\varepsilon} \\ \perp & \text{otherwise} \end{cases}$$

will choose later

How well does this work? Enough to show that

with respect to channel AND code!

$$\underbrace{E}_{\text{average over random choice of code!}}[\underbrace{P_B}_{\text{average over random source message + channel output}}] = \frac{1}{2^K} \sum_s \underbrace{\Pr(\hat{S} \neq s \mid S = s)}_{\text{independent of } s \text{ by symmetry of construction}} \longrightarrow 0$$

Indeed, if true on average for random codes then $\exists$ codes w/ this property!

When is $\hat{S} \neq s$? Recall: $s \longrightarrow X^N(s) \longrightarrow Y^N \longrightarrow \hat{S} = \sigma(Y^N)$.

Two options for errors:

* $(X^N(s), Y^N) \notin J_{N,\varepsilon}$: $\Pr(\cdots) \longrightarrow 0$ by ②

* $(X^N(s'), Y^N) \in J_{N,\varepsilon}$ for some $s' \neq s$:

$K = \lceil NR \rceil \leq N(R + \frac{1}{N})$

$$\Pr(\ldots) \underset{③}{\leq} \#\{s' \neq s\} \cdot 2^{-N(I(X:Y) - 3\varepsilon)} \leq 2^{N(R + \frac{1}{N} - I(X:Y) + 3\varepsilon)}$$

$\longrightarrow 0$ if we choose ε s.th. $R < I(X:Y) - 3\varepsilon$

$\Rightarrow$ $\Pr(\hat{S} \neq s \mid S = s) \longrightarrow 0$ for each s, so also $E[P_B] \longrightarrow 0$ $\square$

Proof of Converse (B)

Two tools:

* Chain rule: $H(AB|C) = H(A|C) + H(B|AC)$

  Pf: $RHS = H(AC) - H(C) + H(ABC) - H(AC) = LHS$ $\square$

* Data Processing Inequality (DPI): If $A \to B \to C$ Markov chain:

  $$I(A:B) \geq I(A:C) \quad \& \quad H(A|B) \leq H(A|C)$$

  i.e. $P(a,b,c) = P(a)P(b|a)P(c|b)$

  → $\boxed{\text{EX CLASS 8}}$

How can we reason about all possible decoders? DPI them away :-)

$\boxed{\text{Fact:}}$ If X^N arbitrary and Y^N channel outputs:

i.e. $P(x^N, y^N) = P(x^N) Q(y_1|x_1) \cdots Q(y_N|x_N)$

$$I(X^N : Y^N) \leq \sum_{i=1}^{N} I(X_i : Y_i) \leq N \cdot C$$

→ $\boxed{\text{HW 5}}$

Fano's inequality: If $S \to Y \to \hat{S}$ Markov chain, $p = \Pr(S \neq \hat{S})$:

$$\underbrace{H(\{p, 1-p\})}_{\leq 1} + p \cdot \log \# \mathcal{A}_S \geq H(S|\hat{S}) \overset{\text{DPI}}{\geq} \underbrace{H(S|Y)}_{\text{indep. of "decoder"!}}$$

WE STOPPED HERE

Pf of Fano: Define $E = \begin{cases} 1 & \hat{S} \neq S \\ 0 & S = S \end{cases}$ s.th. $H(E) = H(\{p, 1-p\})$

Use chain rule in two ways:

$$H(ES|\hat{S}) \overset{\text{chain rule}}{=} H(S|\hat{S}) + \underbrace{H(E|S\hat{S})}_{=0} = H(S|\hat{S}) \overset{\text{DPI}}{\geq} H(S|Y)$$

$$\begin{aligned} H(ES|\hat{S}) &\overset{\text{ditto}}{=} H(E|\hat{S}) + H(S|E\hat{S}) \\ &\leq H(E) + p\, H(S|\hat{S}, E=1) + (1-p) \underbrace{H(S|\hat{S}, E=0)}_{=0 \text{ since } S=\hat{S} \text{ if } E=0} \\ &\leq H(E) + p \cdot \log \# \mathcal{A}_S \end{aligned}$$ $\square$

Now we can prove (B): Let $\frac{K}{N} \geq R > C$ and $S \in \{1, .., 2^K\}$ uniform:

$$H(S|Y^N) = H(S) - I(S:Y^N) \overset{\text{DPI}}{\geq} \underbrace{H(S)}_{=K} - \underbrace{I(X^N:Y^N)}_{\leq N \cdot C \text{ via the } \boxed{\text{Fact}}} \geq K - N \cdot C$$

OTOH: Fano's inequality applied to $S \to Y^N \to \hat{S}$:

$$H(S|Y^N) \leq 1 + p_B \cdot K$$

Together: $K - N \cdot C \leq 1 + p_B \cdot K$

$$\Rightarrow p_B \geq 1 - \frac{NC}{K} - \frac{1}{K} \geq 1 - \frac{C}{R} - \frac{1}{NR} \longrightarrow 1 - \frac{C}{R} > 0$$

Can never go below this error probability for large enough N

$\square$

Are we happy? What questions does Shannon's theorem leave unaddressed? algorithmics, large N, ... how to even compute C ?