Lec 1: Entropy & Mutual Information

Yanjun Han

Entropy. For a discrete RV X taking value in $\mathcal{X}$ with pmf p, its entropy $H(X)$ (or $H(p)$) is defined as

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} .$$

Remarks:

1. $0 \le H(X) \le \log |\mathcal{X}|$ $\left(H(X) \le \log \sum_{x \in \mathcal{X}} p(x) \cdot \frac{1}{p(x)} = \log |\mathcal{X}| \text{ by Jensen} \right)$
2. $H(X)$ can be finite or infinite when $|\mathcal{X}| = \infty$
3. For continuous (or general) RVs, need to find a measure μ s.t. X has a density f w.r.t. μ, and define the differential entropy

$$\underset{\substack{\uparrow \\ \text{usually} \\ \text{lower-case } h}}{h(X)} = \int_{\mathcal{X}} f(x) \log \underset{\substack{\uparrow \\ \text{value depends on the choice of } \mu}}{\frac{1}{f(x)}} \, d\mu(x)$$

4. Base of log: for IT applications (only this lecture), take $\log = \log_2$ (bits);
   For other applications (all later lectures), $\log = \log_e$ (nats).

Why entropy? Shannon (1948) shows that entropy characterizes the fundamental limit of source coding.

Source coding problem (for the i.i.d. case)

Given: ① an input alphabet $\mathcal{X}$ (e.g. all English letters $\{a, b, \cdots, z\}$)
② a known pmf p on $\mathcal{X}$ (i.e. the source distribution)

Target: find a map (i.e. code) $f : \mathcal{X} \to \{0,1\}^* := \bigcup_{n=1}^{\infty} \{0,1\}^n$, such that

① it's uniquely decodable, i.e. based on the concatenation $(f(x_1), \cdots, f(x_m))$, one can uniquely decode m and $(x_1, \cdots, x_m) \in \mathcal{X}^m$

② the expected codelength $\mathbb{E}[\underbrace{\ell}_{\ell(\cdot)\text{: length of the codeword (in bits)}}(f(X))] = \sum_{x \in \mathcal{X}} p(x) \ell(f(x))$ is minimized

Example. If $\mathcal{X} = \{a, b, c\}$ and $p = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$

(a) The code $a \to 0$, $b \to 10$, $c \to 11$ is uniquely decodable,

e.g. 1001011 decodes into babc

(b) The code $a \to 0$, $b \to 1$, $c \to 10$ is NOT uniquely decodable.

e.g. 10 decodes into either c or ba

(c) The code $a \to 10$, $b \to 0$, $c \to 11$ is uniquely decodable and has a smaller expected codelength $2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.5$ bits < 1.75 bits for (a).

Given a length profile $\{\ell_x\}_{x \in \mathcal{X}}$, is there a uniquely decodable code f with $\ell(f(x)) = \ell_x$?

Theorem (Kraft - McMillan)

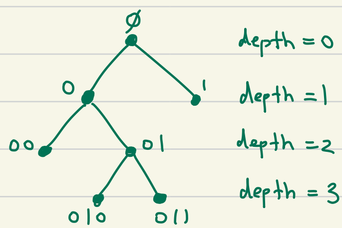
A necessary and sufficient condition is

$$\sum_{x \in \mathcal{X}} 2^{-\ell_x} \le 1. \quad \text{(Kraft inequality)}$$

Pf. (Sufficiency) First note that for a full binary tree (i.e. a binary tree where each node has 0 or 2 children), then

$$\sum_{\text{leaf node } v} 2^{-\text{depth}(v)} = 1.$$

Because $\sum_{x \in \mathcal{X}} 2^{-\ell_x} \le 1$, one can construct a full binary tree s.t. $\mathcal{X} \subseteq \{\text{all leaf nodes}\}$ and $\text{depth}(x) = \ell_x$, $\forall x \in \mathcal{X}$.



A full binary tree, with codewords $\{1, 00, 010, 011\}$

Now use the coding scheme in the example, which results in a prefix code, i.e. no codeword is a prefix of the other. Easy to show that (Exercise) prefix codes are uniquely decodable.

(Necessity) WLOG assume $|\mathcal{X}| < \infty$ and $l_{max} := \max_{x \in \mathcal{X}} l_x < \infty$.

Use a tensor power trick: for uniquely decodable code f.

$$\left(\sum_{x \in \mathcal{X}} 2^{-l(f(x))} \right)^m = \sum_{x_1, \cdots, x_m \in \mathcal{X}} 2^{-(l(f(x_1)) + \cdots + l(f(x_m)))}$$

$$= \sum_{x_1, \cdots, x_m \in \mathcal{X}} 2^{-l(\overbrace{(f(x_1), \cdots, f(x_m))}^{\text{concatenation}})}$$

$$= \sum_{l=1}^{m l_{max}} 2^{-l} \{\# \text{ of concatenated codewords of total length } l\}$$

$$\leq \sum_{l=1}^{m l_{max}} 2^{-l} \cdot 2^{l} \quad (\text{by uniquely decodable assumption})$$

$$= m l_{max}$$

$$\Longrightarrow \sum_{x \in \mathcal{X}} 2^{-l(f(x))} \leq (m l_{max})^{1/m} \xrightarrow{m \to \infty} 1. \qquad \square$$

Using Kraft inequality, we obtain the following characterization of the smallest expected codelength.

> **Thm (Source coding theorem for uniquely decodable code)**
>
> $$H(X) \leq \min_{\substack{\text{uniquely decodable} \\ f}} \mathbb{E}[l(f(X))] < H(X) + 1$$

Pf. (Upper bound) $l_x = \lceil \log_2 \frac{1}{p(x)} \rceil$ satisfies Kraft inequality, and

$$\sum_{x \in \mathcal{X}} p(x) l_x < \sum_{x \in \mathcal{X}} p(x) \left(\log_2 \frac{1}{p(x)} + 1 \right) = H(X) + 1.$$

(Lower bound) Easy to show via Lagrangian multipliers that

$$\left\{ \begin{array}{ll} \min_{l \in \mathbb{R}_+^{|\mathcal{X}|}} & \sum_x p(x) l_x \\ \text{s.t.} & \sum_x 2^{-l_x} \leq 1 \end{array} \right\} = \sum_x p(x) \log_2 \frac{1}{p(x)} = H(X) \qquad \square$$

Remark: 1. The gap between $H(X)$ and $H(X)+1$ could be significant (e.g. when $H(X) = 1.5$ bits). However, in practice, the alphabet $\mathcal{X}$ is usually "super-symbols", e.g. $\mathcal{X} = \{a, \cdots, z\}^{256}$ instead of $\{a, \cdots, z\}$. In such cases, $H(X) \gg 1$ bit.

2. Information theory is usually good at proving "robust" results even if a small error probability can be tolerated; in contrast, the above combinatorial argument fails to do so. See more details below.

Asymptotic equipartition property (AEP).

Another way to write the entropy is

$$H(X) = \mathbb{E}_{X\sim p}\left[\log \frac{1}{p(X)}\right]$$

(★ Warning: Some of you might not be used to seeing the distribution p to appear in BOTH the expectation AND the function)

Therefore, if $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p$, LLN leads to (if $H(X) < \infty$)

$$\frac{1}{n}\log\frac{1}{p(X_1,\cdots,X_n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{p(X_i)} \xrightarrow{\text{a.s.}} \mathbb{E}\left[\log\frac{1}{p(X)}\right] = H(X), \quad \text{as } n\to\infty.$$

$$\Rightarrow \forall \varepsilon > 0, \quad \mathbb{P}\big(\underbrace{p(X_1,\cdots,X_n) \in [2^{-n(H(X)+\varepsilon)}, 2^{-n(H(X)-\varepsilon)}]}_{\text{call this set } T_n^{\varepsilon} \text{ (typical set)}}\big) \longrightarrow 1 \text{ as } n\to\infty.$$

$\Rightarrow$ **Thm (AEP).** The typical set T_n^{ε} satisfies that

① $\mathbb{P}((X_1,\cdots,X_n)\in T_n^{\varepsilon}) \longrightarrow 1$, as $n\to\infty$.

② $(1-o(1))\, 2^{n(H(X)-\varepsilon)} \le |T_n^{\varepsilon}| \le 2^{n(H(X)+\varepsilon)}$

In other words, AEP states that for $X_1, \cdots, X_n \overset{i.i.d.}{\sim} P$, the joint distribution of $X_1, \cdots, X_n$ is "roughly" a uniform distribution over $\doteq 2^{nH(P)}$ typical sequences.

<u>Source coding theorem with error probability</u>.

Diagram: $\boxed{X_1, \cdots, X_n \overset{i.i.d.}{\sim} P} \xrightarrow{\text{encoder}} \boxed{Y \in \{0,1\}^*} \xrightarrow{\text{decoder}} \boxed{(\hat{X}_1, \cdots, \hat{X}_n)}$

with a block error guarantee $\mathbb{P}((X_1, \cdots, X_n) \neq (\hat{X}_1, \cdots, \hat{X}_n)) \le \delta$.

<u>Thm.</u> ① Achievability: $\exists$ (encoder, decoder) s.t. $\frac{1}{n}\mathbb{E}[\ell(Y)] \le H(P) + o(1)$ and $\delta = o(1)$.

② Converse: if $\delta = o(1)$, then ANY (encoder, decoder) pair satisfies $\frac{1}{n}\mathbb{E}[\ell(Y)] \ge H(P) - o(1)$.

<u>Pf.</u> (Achievability) Consider an encoder-decoder pair that enumerates all typical sequences in T_n^ε and ignores all others. Then by AEP,

$$\text{error prob.} = \mathbb{P}((X_1, \cdots, X_n) \notin T_n^\varepsilon) \to 0$$

$$\ell(Y) \le \log_2 |T_n^\varepsilon| \le n(H(P) + \varepsilon) \quad \text{deterministically.}$$

Since $\varepsilon > 0$ is arbitrary, the achievability follows.

(Converse) Fix any $\varepsilon > 0$. Define two sets

$$A = \{(X_1, \cdots, X_n) : \ell(Y) > n(H(P) - 2\varepsilon)\}$$

$$B = \{(X_1, \cdots, X_n) : (X_1, \cdots, X_n) = (\hat{X}_1, \cdots, \hat{X}_n)\}.$$

then

$$\mathbb{P}(T_n^\varepsilon \cap B) \ge 1 - \delta - o(1) \quad \text{by AEP and union bound.}$$

Moreover,

$$|T_n^\varepsilon \cap B \cap A^c| = |\{(x_1,\cdots,x_n) \in T_n^\varepsilon \cap B : \ell(Y(x_1,\cdots,x_n)) \le n(H(P)-2\varepsilon)\}|$$

$$\le |\{y \in \{0,1\}^* : \ell(y) \le n(H(P)-2\varepsilon)\}|$$

by defn. of B, if $(x_1,\cdots,x_n), (x_1',\cdots,x_n') \in B$ are different, one must have $Y(x_1,\cdots,x_n) \neq Y(x_1',\cdots,x_n')$

$$= \sum_{k=1}^{n(H(P)-2\varepsilon)} 2^k < 2\cdot 2^{n(H(P)-2\varepsilon))}$$

$$\Rightarrow \mathbb{P}(T_n^\varepsilon \cap B \cap A^c) \le 2^{-n(H(P)-\varepsilon)} \cdot |T_n^\varepsilon \cap B \cap A^c| < 2\cdot 2^{-n\varepsilon}.$$

by AEP

Therefore, $\mathbb{P}(T_n^\varepsilon \cap A \cap B) \ge 1-\delta - o(1) - 2\cdot 2^{-n\varepsilon} = 1-o(1)$

$$\Rightarrow \frac{1}{n}\mathbb{E}[\ell(Y)] \ge (H(P)-2\varepsilon)\cdot \mathbb{P}(A) \ge (1-o(1))\cdot(H(P)-2\varepsilon)$$

by Markov's inequality. Since $\varepsilon>0$ is arbitrary, the converse follows. ∎

Joint entropy and mutual information

Similar to $H(X) = \mathbb{E}_X[\log \frac{1}{p(X)}]$, can also define

$$H(X,Y) = \mathbb{E}_{X,Y}[\log \frac{1}{p(X,Y)}] \quad \text{(joint entropy)}$$

$$H(Y|X) = \mathbb{E}_{X,Y}[\log \frac{1}{p(Y|X)}]$$

$$= H(X,Y) - H(X) \quad \text{(conditional entropy)}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

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

$$= \mathbb{E}_{X,Y}[\log \frac{p(X,Y)}{p(X)p(Y)}] \quad \text{(mutual information)}$$

Lemma. $I(X;Y) \geq 0$ (non-negativity of mutual info.)

or equivalently, $H(X) \geq H(X|Y)$ (conditioning reduces entropy)

Pf. There is a one-line proof using convexity / KL divergence (next lecture), but let's present a proof using typicality / AEP that will be useful later.

Define

$$T_n^\varepsilon(X) = \left\{ (x^n, y^n) : \left| \frac{1}{n} \sum_{i=1}^n \log_2 \frac{1}{P_X(X_i)} - H(X) \right| \leq \varepsilon \right\}$$

$$T_n^\varepsilon(Y) = \left\{ (x^n, y^n) : \left| \frac{1}{n} \sum_{i=1}^n \log_2 \frac{1}{P_Y(y_i)} - H(Y) \right| \leq \varepsilon \right\}$$

$$T_n^\varepsilon(X,Y) = \left\{ (x^n, y^n) : \left| \frac{1}{n} \sum_{i=1}^n \log_2 \frac{1}{P_{XY}(x,y)} - H(X,Y) \right| \leq \varepsilon \right\}.$$

and joint typical set $\quad T_n^\varepsilon = T_n^\varepsilon(X) \cap T_n^\varepsilon(Y) \cap T_n^\varepsilon(X,Y)$.

For $(X_1, Y_1), \cdots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_{XY}$, LLN + union bound yields

$$\mathbb{P}((X^n, Y^n) \in T_n^\varepsilon) \xrightarrow{n \to \infty} 1,$$

from which one deduces that $\quad |T_n^\varepsilon| \geq (1 - o(1))\, 2^{n(H(X,Y) - \varepsilon)}$.

Next draw $(\tilde{X}_1, \tilde{Y}_1), \cdots, (\tilde{X}_n, \tilde{Y}_n) \overset{\text{i.i.d.}}{\sim} P_X P_Y$. Then

$$\begin{aligned}
1 &\geq \mathbb{P}((\tilde{X}^n, \tilde{Y}^n) \in T_n^\varepsilon) \\
&= \sum_{(x^n, y^n) \in T_n^\varepsilon} \mathbb{P}(\tilde{X}^n = x^n, \tilde{Y}^n = y^n) \\
&\geq (1 - o(1))\, 2^{n(H(X,Y) - \varepsilon)} \cdot 2^{-n(H(X) + \varepsilon)} \cdot 2^{-n(H(Y) + \varepsilon)} \\
&= (1 - o(1))\, 2^{-n(I(X;Y) + 3\varepsilon)}
\end{aligned}$$

$\Longrightarrow \; I(X;Y) + 3\varepsilon \geq 0$, and $I(X;Y) \geq 0$ by taking $\varepsilon \to 0^+$. $\square$

This is a fundamental inequality to prove other inequalities. e.g.

① $H(X_1, \cdots, X_n) = \sum_{k=1}^{n} H(X_k | X^{k-1}) \leq \sum_{k=1}^{n} H(X_k)$

② If $P_{Y^n|X^n} = \prod_i P_{Y_i|X_i}$, then

$$
\begin{aligned}
I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \qquad \left(H(Y^n|X^n) = \mathbb{E}\left[\log \frac{1}{p(Y^n|X^n)}\right] = \mathbb{E}\left[\sum_i \log \frac{1}{p(Y_i|X_i)}\right] \right) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i; Y_i)
\end{aligned}
$$

③ If $P_{X^n} = \prod_i P_{X_i}$, then

$$
\begin{aligned}
I(X^n; Y^n) &= H(X^n) - H(X^n|Y^n) \\
&= \sum_i H(X_i) - H(X^n|Y^n) \\
&\geq \sum_i H(X_i) - \sum_i H(X_i|Y^n) \qquad (\text{by } ①) \\
&\geq \sum_i H(X_i) - \sum_i H(X_i|Y_i) \qquad (\text{conditioning reduces entropy}) \\
&= \sum_i I(X_i; Y_i)
\end{aligned}
$$

Remark: All inequalities that can be shown via

① monotonicity: $H(X) \leq H(X, Y)$

② submodularity: $H(X_A) + H(X_B) \geq H(X_{A \cup B}) + H(X_{A \cap B})$

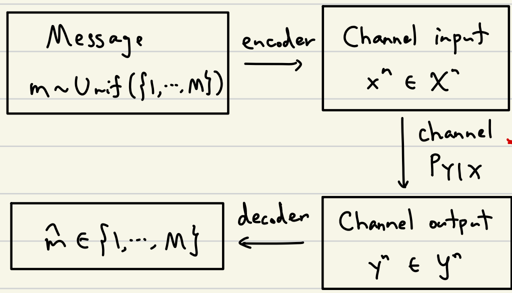
are called <u>Shannon-type inequalities</u>.

Why mutual information? Shannon (1948) shows that it characterizes the fundamental limit of <u>communications/channel coding</u> and <u>lossy compression</u>

(later this lecture) (skipped; related to "mutual info method" for statistics later)

Channel coding problem.

Diagram:



known and given by nature

n channel uses are indepedent

i.e. $P_{Y^n|X^n} = \prod_i P_{Y_i|X_i}$

Goal: Given a (block) error probability guarantee $\mathbb{P}(m \neq \hat{m}) \leq \delta$, aim to send as many messages as possible, or equivalently, maximize the rate of communication

$$R_n = \frac{\log M}{n} \quad \text{(bits per channel use)}$$

Defn (channel capacity).

$$C = C(P_{Y|X}) = \max_{P_X} I(X;Y), \quad \text{with } P_{XY} = P_X P_{Y|X}$$

(In other words, given the transition probability $P_{Y|X}$ from X to Y, design an input distribution P_X s.t. $I(X;Y)$ is maximized)

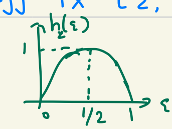
Examples. ① Binary symmetric channel (BSC):

$$P_{Y|X}: \quad \begin{array}{cc} & \quad Y \\ & \begin{array}{cc} 0 & 1 \end{array} \\ X \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix} \end{array}$$

$$I(X;Y) = H(Y) - H(Y|X) \leq 1 - h_2(\varepsilon), \quad \text{with equality iff } P_X = [\tfrac{1}{2}, \tfrac{1}{2}].$$

binary entropy function

$$h_2(\varepsilon) = \varepsilon \log_2 \frac{1}{\varepsilon} + (1-\varepsilon)\log_2 \frac{1}{1-\varepsilon}.$$



② Binary erasure channel (BEC):

$$P_{Y|X}: \quad \begin{array}{cc} & \quad Y \\ & \begin{array}{ccc} 0 & 1 & \perp \end{array} \\ X \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{bmatrix} \end{array}$$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \mathbb{P}(Y \neq \perp)\underbrace{H(X|Y\neq\perp)}_{=0} - \mathbb{P}(Y=\perp)\underbrace{H(X|Y=\perp)}_{=H(X)} \\ &= (1-\varepsilon)H(X) \le 1-\varepsilon, \text{ with equality iff } P_X = [\tfrac{1}{2}, \tfrac{1}{2}]. \end{aligned}$$

Thm (Shannon's channel coding theorem) Fix any $\varepsilon > 0$.

① Achievability: if $R_n < C - \varepsilon$, then $\exists$ (encoder, decoder) s.t.

$$\mathbb{P}(m \neq \hat{m}) \longrightarrow 0 \text{ as } n \to \infty.$$

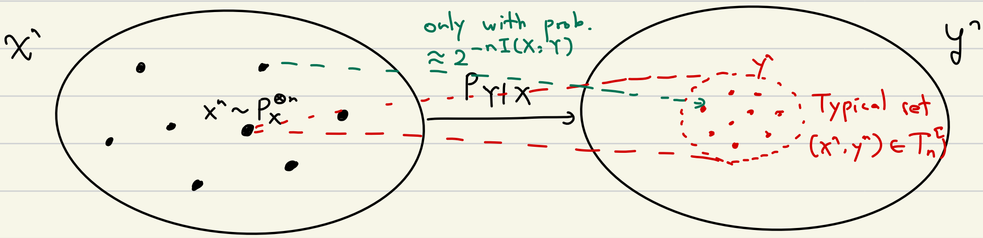
② (Weak) converse: if $R_n > C + \varepsilon$, then $\forall$ (encoder, decoder),

$$\liminf_{n\to\infty} \mathbb{P}(m \neq \hat{m}) > 0.$$

(Strong converse: $\liminf_{n\to\infty} \mathbb{P}(m \neq \hat{m}) = 1$; see Lec 4)

In other words, the maximum rate of communication is (asymptotically) the channel capacity!

Achievability: random coding & typicality.



Random codebook: generate $X^n_{(1)}, \cdots, X^n_{(M)} \overset{\text{i.i.d.}}{\sim} P_X^{\otimes n}$

Encoder: for message $m \in [M]$, send $X^n_{(m)}$

Decoder: find the unique $\hat{m} \in [M]$ s.t. $(X^n_{(\hat{m})}, Y^n)$ is joint typical (see defn. on page 8); if none or not unique, report failure.

Analysis: WLOG assume that the true message is $m=1$.

Then $\hat{m} = m$ if:

① $(X^n_{(1)}, Y^n)$ is joint typical;

② none of $(X^n_{(2)}, Y^n), \cdots, (X^n_{(M)}, Y^n)$ is joint typical.

By LLN, $\mathbb{P}(①) = 1 - o(1)$.

Reversing the analysis on Page 8, since $(X^n_{(2)}, Y^n) \sim P_X^{\otimes n} \otimes P_Y^{\otimes n}$ (independent!!),

$$\mathbb{P}((X^n_{(2)}, Y^n) \text{ joint typical}) \le 2^{-n(I(X;Y)-3\varepsilon)},$$

so union bound gives $\mathbb{P}(②) \ge 1 - M \cdot 2^{-n(I(X;Y)-3\varepsilon)}$. If $\log_2 M < n(I(X;Y)-4\varepsilon)$, then

$$\mathbb{P}(②) \ge 1 - e^{-n\varepsilon} = 1 - o(1).$$

Therefore, $\mathbb{P}(\hat{m} = 1) \ge \mathbb{P}(① \text{ and } ②) = 1 - o(1)$. ☐

Remark: 1. Random coding was a remarkable idea at the time, when algebraic codes were more popular. This also motivated the entire field of probabilistic methods.

2. This coding scheme is computationally expensive. First efficient codes which attains the Shannon limit were found in 2000's, including the spatially coupled LDPC code and polar code.

Weak converse: Fano's inequality.

(How IT-based ideas are robust to errors)

Lemma. (Data processing inequality for MI)

If $X - Y - Z$ forms a Markov chain (i.e. $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$), then

$$I(X;Y) \geq I(X;Z)$$

Pf. Shannon-type inequalities:

$$\begin{aligned} I(X;Y) - I(X;Z) &= H(X|Z) - H(X|Y) \\ &= H(X|Z) - H(X|Y,Z) \quad \text{(By Markov)} \\ &= I(X;Y|Z) \geq 0. \end{aligned}$$

$\boxtimes$

Thm (Fano's inequality) If $X \sim \text{Unif}([M])$.

Then

$$\mathbb{P}(X \neq Y) \geq 1 - \frac{I(X;Y) + \log 2}{\log M}.$$

Pf. Let $E = 1(X \neq Y)$. Then

$$\begin{aligned} H(X|Y) &= H(X|Y,E) + \underbrace{I(X;E|Y)}_{\leq H(E) \leq \log 2} \\ &\leq \mathbb{P}(E=1)\underbrace{H(X|Y,E=1)}_{\leq H(X) = \log M} + \mathbb{P}(E=0)\underbrace{H(X|Y,E=0)}_{=0} + \log 2 \\ &\leq \mathbb{P}(X \neq Y) \cdot \log M + \log 2. \end{aligned}$$

On the other hand,

$$\begin{aligned} H(X|Y) &= H(X) - I(X;Y) \\ &= \log M - I(X;Y), \quad (X \sim \text{Unif}[M]) \end{aligned}$$

rearranging yields the claim. $\boxtimes$

(In Lec 2, we'll see more "principled" proofs of Fano's inequality)

To apply Fano's inequality, if $R_n > C + \varepsilon$,

$$\mathbb{P}(m \neq \hat{m}) \geq 1 - \frac{I(m;\hat{m}) + \log 2}{\log M}$$

$$\geq 1 - \frac{I(X^n; Y^n) + \log 2}{\log M} \quad \text{(Markov structure } m - X^n - Y^n - \hat{m}\text{)}$$

$$\geq 1 - \frac{\sum_{i=1}^{n} I(X_i; Y_i) + \log 2}{\log M} \quad \text{(inequality (2) on Page 9)}$$

$$\geq 1 - \frac{nC + \log 2}{\log M} \quad \text{(defn. of } C\text{)}$$

$$\geq 1 - \frac{nC + \log 2}{n(C+\varepsilon)} \xrightarrow{n \to \infty} \frac{\varepsilon}{C + \varepsilon} > 0,$$

establishing the weak converse. $\boxed{\text{Q.E.D.}}$