8.372 Lecture 3: Classical Info. Theory — 09/17/2026

Overview
I. Intro + Motivation
II. Entropy + noiseless coding
III. Conditional entropy + mutual info.
IV. Noisy channel coding theorem

I. Intro + motivation

- Classical Inf. theory is about fundamental limits of inf. processing tasks — e.g., sending, storing or transmitting/learning info. — in the asymptotic setting.
  ↳ many uses/copies of some resource, some small error prob. going to zero.
- Story becomes more complicated in the quantum setting, but the most basic theorem statements carry over in some way from the classical cases. In particular, we can summarize some of these in the following table.

Application	Classical	Quantum
Data compr.	Shannon entropy	von Neumann entropy
Channel coding	Mutual info.	Quantum mutual info. (for cl. info. over q. channel) or coherent info. (for q. info. over q. channel)
Hypothesis testing	Relative entropy	Quantum relative entropy

(this lecture)

II. Entropy + noiseless source coding

- entropy is a measure of uncertainty; or the expected amount of surprise upon learning the value of a random variable.

Def^n: Let $p: \mathcal{X} \to [0,1]$ be a distr. over some finite alphabet $\mathcal{X}$, let $X \sim p$ be a random variable. The **Shannon entropy** of X is

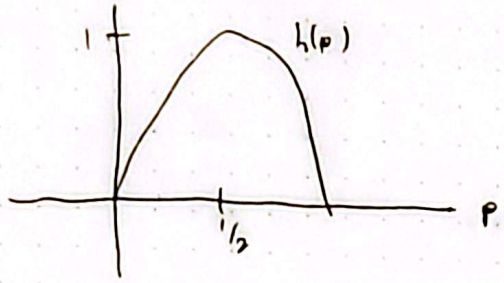
$$H(X) \equiv H(p) \equiv -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \qquad 0 \log 0 \equiv 0.$$

- Some properties:

1. (Bounds) $0 \le H(X) \le \log_2 d$ for $d = |\mathcal{X}|$
   (0: deterministic; $\log_2 d$: unif.)

2. (Concavity) $\forall\ 0 \le \lambda \le 1$ we have
$$H(\lambda p_1 + (1-\lambda) p_2) \ge \lambda H(p_1) + (1-\lambda) H(p_2).$$

- Can plot binary entropy $h(p) \equiv -p\log p - (1-p)\log(1-p)$.



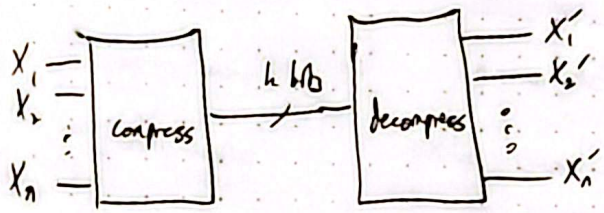
- For multiple random variables $p: \mathcal{X}\times\mathcal{Y} \to [0,1]$, $(X,Y)\sim p$, nothing changes:
$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y).$$

- Note: If p is a product distr. $H(X,Y) = H(X)+H(Y)$.

- Noiseless source coding setup:

$X\sim p$

$X_i \sim p$ iid



- We say a rate R is achievable if the message can be recovered w/ vanishing prob. of error as $n\to\infty$ when $k = n(R+\delta)$ for any $\delta > 0$.

Theorem (noiseless source coding): The optimal rate of compression $R^* := \inf\{R : R \text{ is achievable}\}$ is given by $H(X)$, i.e., $R^* = H(X)$.

Pf sketch: Consider the distr. $p^n: \mathcal{X}^n \to [0,1]$ over n-letter words, & suppose w.l.o.g. $\mathcal{X} = \{1,2,\dots,d\}$. Since letters are iid,
$$p^n(x_1,x_2\dots x_n) = \prod_{i=1}^{n} p(x_i).$$
typically
By LLN, any letter $\ell\in[d]$ occurs $\approx np(\ell)$ times, as $n\to\infty$.
$\therefore$ by Stirling, # of typical strings is $\dfrac{n!}{\prod_{\ell=1}^{d}(np(\ell))!} \approx 2^{nH(X)}$.

Why?

$$\underbrace{11\dots1}_{np(1)}\underbrace{22\dots2}_{np(2)}\dots\underbrace{dd\dots d}_{np(d)}$$

and all distinct strings formed by permuting elements are the typical strings.

III. Conditional Entropy & Mutual Info.

Def'n: For random variables $(X, Y) \sim p$ we call the **conditional entropy of Y given X** is

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= H(X,Y) - H(X). \end{aligned}$$

→ $H(Y|X)$ is the uncertainty of Y once we know X. Think of the case where X & Y are perfectly correlated, for example. Then it is straightforward to see that $H(Y|X) = 0$.

Mutual Info: The **mutual information** between X and Y is

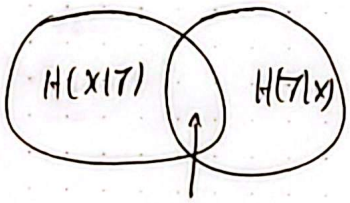
$$\begin{aligned} I(X:Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \qquad \text{* note the symmetry in } X \text{ & } Y \end{aligned}$$

- Remark: mutual info. is nonnegative, both classically and quantumly.

$$\begin{aligned} I(X:Y) \geq 0 &\iff H(X) + H(Y) \geq H(X,Y) \\ &\iff H(Y|X) \leq H(Y). \end{aligned}$$

→ we'll see why later.

- Picture:

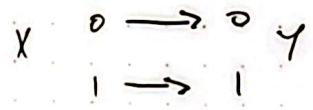


- Information channel capacity.
  - Let N be a noisy channel, $X \to \boxed{N} \to Y$, which specifies $p(y|x)$.
  - The "information channel capacity" of N is defined as

$$C = \max_{p(x)} I(X:Y).$$

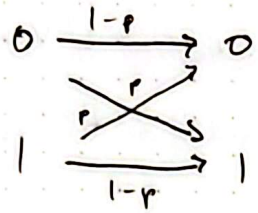
Examples

noiseless



$$\max_X I(X;Y) = \max_X [H(X) - H(X|Y)]$$
$$= \boxed{1}.$$

Binary symmetric channel

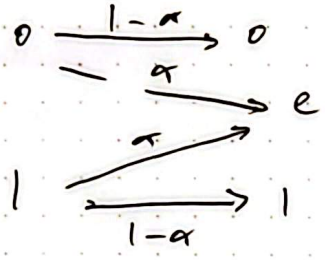


$$I[X:Y) \triangleq H(Y) - \sum_x p(x) H(Y|X=x)$$
$$\overset{\triangle}{=} H(Y) - H(p)$$
$$\leq 1 - h(p).$$

If we take X unif., equality attained

$$\Rightarrow I(X;Y) = \boxed{1 - h(p)}.$$

Erasure

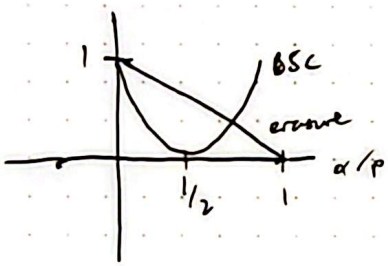


$$C = \max_x H(Y) - H(Y|X)$$
$$= \max_{p(x)} H(Y) - H(\alpha)$$

Let E denote r.v. indicating if erasure occurred.

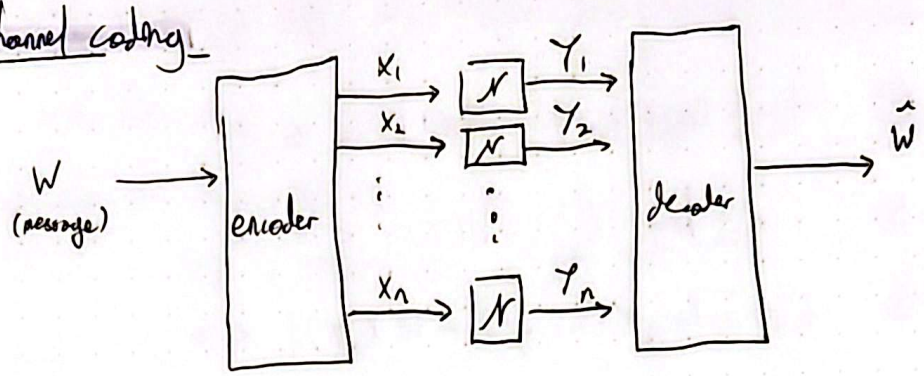
$$H(Y) = H(E) + H(Y|E)$$
$$= h(\alpha) + (1-\alpha) H(p), \quad p \equiv P_1[X=1]$$

$$\therefore C = \max_p (1-\alpha) h(p) + h(\alpha) - h(\alpha)$$
$$= \boxed{1-\alpha}.$$



IV. Noisy channel coding

- Setup is



- We say this is an (M, n) code if the message W can take on M possible values. We say the code has rate R if $M = 2^{nR}$, and R is achievable if prob. error in decoding i.e., $P_r[\hat{w} \neq w]$, goes to zero as $n \to \infty$

Theorem (noisy channel coding): The optimal rate $R^* := \sup\{R : R \text{ is achievable}\}$, known as the channel capacity of $\mathcal{N}$, is equal to the information channel capacity of $\mathcal{N}$. I.e.,

↳ (achievability only) $R^* = C = \max_X I(X:Y)$.

Pf. Generate a $(2^{nR}, n)$ code at random according to some fixed $p(x)$. The code is

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

Each entry is generated iid according to $p(x)$. Thus, prob. that we generate a particular code is $P_r[C] = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w))$.

Receiver gets Y^n according to

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

Strategy: use jointly typical decoding:

→ declare "$\hat{w}$ sent" if $(X^n(\hat{w}), Y^n)$ is jointly typical, & there is no other index w' s.t. $(X^n(w'), Y^n)$ is jointly typical.

But first: what is joint typicality? →

↳ "typical set", "jointly typical set"

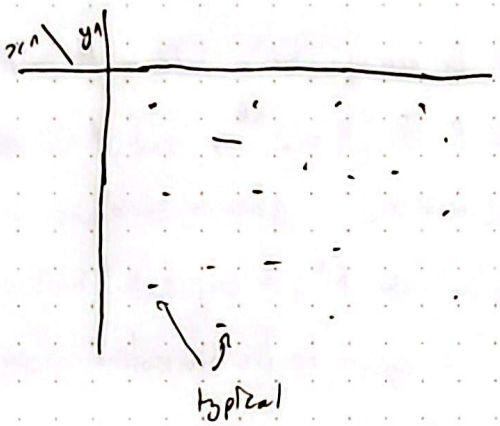
Set $A_\epsilon^{(n)} := \{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } ①, ②, ③ \text{ hold} \}$

where ① $\left| -\frac{1}{n} \log p^n(x^n) - H(X) \right| < \epsilon$

② $\left| -\frac{1}{n} \log p^n(y^n) - H(Y) \right| < \epsilon$

③ $\left| -\frac{1}{n} \log p^n(x^n, y^n) - H(X,Y) \right| < \epsilon$.

Intuition:



Most pairs (x^n, y^n) are not typical. But if $(X^n, Y^n) \sim p$, then typicality is overwhelmingly likely.

Next we will bound the average prob. of error, averaged over randomness due to (i) choice of code C, (ii) choice of message W, & (iii) ~~choice of~~ randomness due to N.

→ if error is low, we can convert (ii) (average case statement) to worst case statement via ~~probabilistic existence argument~~ discarding some messages.

→ ~~If~~ (i) can be converted to statement of existence of some fixed code C that succeeds, via probabilistic existence.

$$P_r[\text{error}] = \sum_C P_r[C] \cdot (\text{avg. error w/ code } C)$$

$$= \sum_C P_r[C] \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C)$$

↳ error on input w w/ code C

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C P_r[C] \lambda_w(C)$$

no w-dependence, by symmetry of C randomness.

$$= P_r[\text{error} \mid W=1].$$

$$\therefore P_r[\text{error} \mid W=1] \le P_r[(X^n(1), Y^n) \notin A_\epsilon^{(n)}] + \sum_{i=2}^{2^{nR}} P_r[(X^n(i), Y^n) \in A_\epsilon^{(n)}].$$

(A) (B)

For Ⓐ, $P_r[(X^n, Y^n) \in A_\epsilon^{(n)}] \to 1$ as $n \to \infty$ is just the LLN.

For Ⓑ, note that since $X^n(2)$ indep. of $X^n(1)$, $X^n(2), Y^n$ are indep. as well, & so on for $X^n(3)$, $X^n(4)$, etc.

$$\therefore \quad P_r[(X^n(2), Y^n) \in A_\epsilon^{(n)}] = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n)\, p(y^n)$$

joint typicality →

$$\leq \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} 2^{-n[H(X)-\epsilon]}\, 2^{-n[H(Y)-\epsilon]}$$

$$\leq 2^{n[H(X,Y)+\epsilon]}\, 2^{-n[H(X)-\epsilon]}\, 2^{-n[H(Y)-\epsilon]}$$

$$= 2^{-n[I(X:Y)-3\epsilon]}$$

so overall Ⓑ $\leq 2^{nR} \cdot 2^{-n[I(X:Y)-3\epsilon]}$

$$\approx 2^{-n[I(X:Y)-R]}$$

which is small so long as $R < I(X:Y)$. Take the sup over $p(x)$. This proves achievability.