Lecture 3 (Jan 22)

Exercise. $H(X|Y)=0 \Leftrightarrow X=g(Y)$ for some func. $g(\cdot)$

Chain rule for MI

$$I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1})$$

Proof:

$$\text{LHS} = H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y)$$

$$= \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1},Y) = \checkmark$$

M.I. under conditioning.

$I(X;Y)$ vs. $I(X;Y|Z)$?

① if $Z=Y \Rightarrow I(X;Y|Z)=0 \le I(X;Y)$.

② $Z=X\oplus Y$ & $X\perp Y$ $\Rightarrow I(X;Y)=0 < I(X;Y|Z)=1$.

$$H(X_1, \ldots, X_n) \le H(X_1) + \cdots + H(X_n)$$

& $=$ if X_i are independent.

Lemma: If $X_1, \ldots, X_n$ are independent,

$$I(X_1, \ldots, X_n ; Y) \ge \sum_{i=1}^{n} I(X_i ; Y)$$

LHS. $= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n \mid Y)$

$$\overset{\text{indep}}{=} \sum_{i=1}^{n} H(X_i) \cancel{+ H(Y)} -$$

$$\left(\cancel{H(Y)} + H(X_1 \mid Y) + H(X_2 \mid X_1, Y) + \cdots + H(X_n \mid X_1, \ldots, X_{n-1}, Y)\right)$$

$$\overset{\text{regroup}}{=} \sum_{i=1}^{n} H(X_i) - H\underbrace{(X_i \mid X_1, \ldots, X_{n-1}, Y)}_{\le H(X_i \mid Y)}$$

$$\ge \sum_{i=1}^{n} I(X_i ; Y).$$

Def.

Relative Entropy / Information Divergence / Kullback-Liebler Divergence (KL) :

Let p, q be distributions on $\mathcal{U}$, Then,

$$D(p \| q) \overset{\text{def}}{:=} \sum_{x \in \mathcal{U}} p(x) \log \frac{p(x)}{q(x)}.$$

⚠ Not symmetric!

* Clearly, for $p = q \Rightarrow D(p\|q) = 0$.

* $D(p\|q)$ finite $\Rightarrow$ $\text{Supp}(p) \subseteq \text{Supp}(q)$.

* Gibbs inequality: $D(p\|q) \geq 0$ & $= 0$ iff $p = q$.

Proof. (Concavity)

$\left[Z = \frac{q(x)}{p(x)} \text{ w.p. } p(x) \right]$

$$-D(p\|q) = \sum_x p(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_x \log Z$$

$$\overset{\text{Jensen}}{\leq} \log \mathbb{E}(Z) = \log 1 = 0.$$

(equality iff $Z = \mathbb{E}(Z)$ everywhere)

Theorem. $I(X;Y) = \mathcal{D}\left(p(x,y) \,\|\, p(x)p(y) \right)$

$\begin{cases} \text{in particular, if } X \perp Y \Rightarrow \mathcal{D}(\cdot \| \cdot) = 0. \checkmark \\ \text{on the other hand: if } X = Y \Rightarrow \mathcal{D} \text{ is maximal.} \end{cases}$

Proof.

RHS. $= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= \underbrace{\sum_{x,y} p(x,y) \log \frac{1}{p(x)}}_{H(X)} - \underbrace{\sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}}_{H(X|Y)}$$

$$= I(X;Y).$$

$\square$

Some viewpoints ~~of~~ on KL-divergence.

① Source Coding interpretations:

"Increase in expected encoding length due to incorrect knowledge of the distribution"

② Rejection Sampling

$p, q \leftarrow$ Universe $\mathcal{U}$.

Setting: Sequence $X_1, X_2, \ldots$ iid $\leftarrow q$.

Goal: Output i^* s.t. X_{i^*} is distributed according to p.
(with no modification)

* Say p is uniform on $\{1, \ldots 50\}$
& q " " " $\{1, \ldots 100\}$.

Then, Sample from q until you hit $\{1, \ldots 50\}$.
but the inverse is not possible.

Solution: Output the first i^* s.t. $X_{i^*} \leq 50$

$\Rightarrow \mathbb{E}(i^*) = 2, \; \mathbb{E}(\text{length } i) = 1.$

Thm. $\exists$ strategy that achieves

$$\mathbb{E}(\text{length } i^*) \lesssim D(p \| q).$$

over p & internal randomness of strategy.

(No proof now!)

③. (Compressing protocols)

knows p Ⓐ

knows q Ⓑ

$\downarrow$

Goal: Communicate $X \leftarrow p$ to B.

($\exists$ common shared randomness)

<u>Thm.</u> $\exists$ interactive protocol between A & B with Expected "$\mathcal{D}(p \| q)$ + small" bits of communication st. in the end,

① A outputs ~~answer~~ $a \leftarrow p$

② B outputs b st.

$$\forall x, \ \Pr(b = x \mid a = x) \geq 1 - \varepsilon.$$

(i.e, $\mathcal{D}(\|)$ is the complexity of "agreement").

---

② & ③ are useful for Compression of protocols.

Data processing inequality:

Suppose: X, Y, $g(Y)$ deterministic func.

Then:

$$I(X;Y) \geq I(X; g(Y)).$$

Proof next time!