CO781 / QIC 890 Lec 08, Oct 4, 2016

Topic 3: classical communication via classical channels

Let X, Y be two rv's with joint distribution $p(xy)$

and sample space $\Omega_X \times \Omega_Y$

① $H(XY) = -\sum_{xy} p(xy) \log p(xy)$ (as defined earlier)

② $\forall y \in \Omega_Y$, let $p(y) = \sum_x p(xy)$ the marginal dist^n on Y.

The conditional distribution of X given Y=y :

$$q_y(x) = \frac{p(xy)}{p(y)} \quad \text{if } p(y) \neq 0$$

Def [Conditional entropy]

Using the above notations, the entropy of X conditioned on Y is :

$$H(X|Y) := \underbrace{\sum_y p(y) \underbrace{H(q_y)}_{\text{entropy of X given } Y=y}}_{\text{Averaged over } Y}$$

Thm (Chain rule)

$H(XY) = H(Y) + H(X|Y)$.

Pf: $H(XY) = -\sum_x \sum_y p(xy) \log \underbrace{p(xy)}_{p(y)\, q_y(x)}$

$$= -\sum_y \underbrace{\sum_x p(xy)}_{p(y)} \log p(y) - \underbrace{\sum_y \sum_x p(y)\, q_y(x) \log q_y(x)}_{H(q_y)}$$

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

Cor: $H(XY) = H(YX)$, $H(XY) = H(X) + H(Y|X)$

Def: [Relative entropy] [Kullback Leibler distance]

Let $p(x)$, $q(x)$ be two distributions on Ω.

$$D(p\|q) := \sum_{x\in\Omega} p(x) \log \frac{p(x)}{q(x)}$$

NB if $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$ then $D(p\|q) = \infty$, $0 \log \frac{0}{q} = 0$, $p \log \frac{p}{0} = \infty$.

NB: D non negative, $D=0 \Leftrightarrow p=q$, but not sym, no Δ ineq (not metric).

* Not used in our course:

- $D(p\|q)$ measures the inefficiency of assuming that the dist^n is q when it's p.
- In hypothesis testing, dist^n is either p or q and n iid samples are given.

Any algorithm to discriminate p, q, WLOG is deterministic, and

output $\begin{cases} p & \text{if } x^n \in A \\ q & \text{" " } \notin \text{"} \end{cases}$ for some $A \subseteq \Omega^n$.

Subject to the constraint prob(output $p \mid p$) $\geq 1-\varepsilon$

let $\beta(n,\varepsilon) := \min\limits_{\substack{A \subseteq \Omega^n \\ p(A) \geq 1-\varepsilon \text{ under } p^n}} q^n(A) \sim 2^{-nD(p\|q)}$.

Pf similar to AEP.

Def [Mutual Info]

$$I(X;Y) := D(p(xy) \,\|\, p(x)\cdot p(y))$$

$$= \sum_{xy} p(xy) \log \frac{p(xy)}{p(x)\cdot p(y)}$$

Properties of H, D, & I: (Cover & Thomas Cp 2)

(H1) Range: $0 \leq H(X) \leq \log|\Omega|$

"=" iff $\exists a$ s.t $p(x)=0 \ \forall x \neq a$

"=" iff $p(x) = \frac{1}{|\Omega|} \ \forall x$

(H2) $H(X|Y) \geq 0$ (from def & (H1), conv combination of entropies)

(H3) $H(XY) \geq H(X)$ (from (H2) & chain rule). The more the merrier / messier

(D4) $D(p\|q) \geq 0$.

Pf: We need Jensen's ineq, that $\mathbb{E} f(X) \geq f(\mathbb{E}(X))$ for any convex function f & rv X, with "=" iff X const.

Then: $$-D(p\|q) = -\sum_{x \in \mathrm{Supp}(p)} p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_{x \in \mathrm{Supp}(p)} p(x) \log \frac{q(x)}{p(x)}$$

$$\overset{JI}{\leq} \log \sum_{x \in \mathrm{Supp}(p)} p(x) \cdot \frac{q(x)}{p(x)}$$

$$= 0$$

Since $\log(\cdot)$ is strictly concave, $D(p\|q) = 0$ iff $\frac{p}{q} = $ const

iff $p(x) = q(x) \ \forall x$

(I5) $I(X:Y) \geq 0$, "=" iff X, Y independent.

(I6) $I(X:Y) = H(X) - H(X|Y)$

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

$\underset{\text{Chain rule}}{=} H(X) + H(Y) - H(XY)$

Interpretation: $I(X:Y)$ = decrease in ignorance of X given Y

= amount of info of X carried by Y.

Note: (I6) follows from the defs of H & I

but it is NOT a definition.

Pf: $I(X:Y) = \sum_{xy} p(xy) \log \frac{p(xy)}{p(x)p(y)}$

$= \sum_{xy} p(xy) \log \frac{q_y(x)}{p(x)}$

$= \sum_{x} \underbrace{\left(\sum_y p(xy)\right)}_{p(x)} \cdot \log \frac{1}{p(x)} + \underbrace{\sum_{xy} p(y) \cdot q_y(x) \log q_y(x)}_{-H(q_y(x))}$

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

(H7) Subadditivity (SA)

$H(XY) \leq H(X) + H(Y)$, "=" iff X, Y indep.

Pf: follows from (I5) & (I6)

(H8) Conditioning reduces entropy

$H(X|Y) \leq H(X)$, "=" iff X, Y indep.

Pf: from (I5) & (I6)

(H9) For 3 rv's X, Y, Z, $H(X|Z) \geq H(X|YZ)$.

Pf: Let $p(xyz)$ be the joint dist^n for XYZ.

For each z, let $q_z(xy) = \dfrac{p(xyz)}{p(z)}$

where $p(z) = \sum_{xy} p(xyz)$.

Let $\hat{X}\hat{Y}$ has distribution $q_z(xy)$.

We have $H(\hat{X}) \geq H(\hat{X}|\hat{Y})$ from previous discussion (*)

- $H(\hat{X}) =$ Entropy on $\eta_z(x) = \sum_y q_z(xy)$

But $\eta_z(x) = \sum_y \dfrac{p(xyz)}{p(z)} = \dfrac{p(xz)}{p(z)} = p_z(x)$ from $\sum_y p(xyz)$

$\therefore H(\hat{X}) = H(X|Z=z)$ evaluated on $\sum_y p(xyz)$.

- $H(\hat{X}|\hat{Y}) = \sum_y \left(\sum_x q_z(xy)\right) \cdot H(\hat{X}|\hat{Y}=y)$

$H(\hat{X}|\hat{Y}=y) =$ Entropy on $\dfrac{q_z(xy)}{\sum_x q_z(xy)}$

$= \dfrac{\frac{p(xyz)}{p(z)}}{\sum_x \frac{p(xyz)}{p(z)}} = \dfrac{p(xyz)}{\sum_x p(xyz)} =$ dist on x given $Y=y$, $Z=z$ evaluated on $p(xyz)$

$\sum_x q_z(xy) = \sum_x \dfrac{p(xyz)}{p(z)} = \dfrac{p(yz)}{p(z)} =$ prob of y given $Z=z$.

$\therefore H(\hat{X}|\hat{Y}) = H(X|Y\,Z=z)$.

- Now take convex combination of the ineq (*) over z

$$\sum_z p(z)\, H(\hat{X}) \geq \sum_z p(z)\, H(\hat{X}|\hat{Y})$$

$$\begin{array}{ccc} \| & & \| \\ H(X|Z) & & H(X|YZ) \end{array}$$

(I10) Def: The conditional mutual information

$$I(X:Y|Z) := H(X|Z) - H(X|YZ)$$
$$= H(Y|Z) - H(Y|XZ)$$

By (H9), $I(X:Y|Z) \geq 0$

(H10) Concavity: mixing increases entropy

Let $\alpha_k \geq 0$, $\sum_{k=1}^{t} \alpha_k = 1$, $p_1, \ldots, p_t$ distributions on Ω

Then $\sum_{k=1}^{t} \alpha_k H(p_k) \leq H\left(\sum_{k=1}^{t} \alpha_k p_k\right)$

Pf: let KX be rv's on $\{1,2,\ldots,t\} \times \Omega$

$\text{prob}(kx) = \alpha_k \, p_k(x)$

Then $\sum_{k=1}^{t} \alpha_k H(p_k) = H(X|K)$

$H\left(\sum_{k=1}^{t} \alpha_k p_k\right) = H(X)$

$\therefore$ (H9) follows from (H8)

(H11) Strong subadditivity SSA or data processing inequality DPI

• Def: We say $X \to Y \to Z$ (a markov Chain)

if $p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$

(ie if $p(z|y) = p(z|xy)$)

• DPI: If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Pf: $I(X;YZ)$

$= \underbrace{H(X) + H(Z) - H(XZ)} + H(XZ) - H(Z) - (H(XYZ) - H(YZ))$

$= I(X:Z) + H(X|Z) - H(X|YZ) \underset{(H9)}{\geq} I(X:Z)$

$I(X;YZ)$

$= \underbrace{H(X) + H(Y) - H(XY)} + H(XY) - H(Y) - [H(XYZ) - H(YZ)]$

$= I(X:Y) + H(Z|Y) - H(Z|XY) = I(X:Y)$

since $p(z|y) = p(z|xy)$

$\therefore I(X:Y) \geq I(X:Z)$

- In particular, if we process Y to obtain Z for any X that may be correlated with Y, we have $p(z|y) = p(z|xy)$.

So such processing can increase mutual info (from $I(X:Y)$ to $I(X:Z)$).

Def [Jointly typical sequence]

Given a distribution $p(xy)$, drawn iid n times

$x^n y^n$ is δ- jointly typical if

ⓐ $\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \delta$ (x^n typical)

ⓑ $\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| \leq \delta$ (y^n typical)

ⓒ $\left| -\frac{1}{n} \log p(x^n y^n) - H(XY) \right| \leq \delta$ ($(xy)^n$ typical)

Def [Jointly typical set]

$A_{n,\delta} = \{ x^n y^n \in \Omega_X^n \times \Omega_Y^n : x^n y^n \text{ jointly typical} \}$

Obs: if $x^n y^n \in A_{n,\delta}$,

then $p(x^n | y^n) = \dfrac{p(x^n y^n)}{p(y^n)} \leq \dfrac{2^{-n(H(XY)-\delta)}}{2^{-n(H(Y)+\delta)}} \leq 2^{-n(H(X|Y)-2\delta)}$

$p(x^n | y^n) = \dfrac{p(x^n y^n)}{p(y^n)} \geq \dfrac{2^{-n(H(XY)+\delta)}}{2^{-n(H(Y)-\delta)}} \geq 2^{-n(H(X|Y)+2\delta)}$

NB. For this y^n, there are $\approx 2^{nH(X|Y)}$ $\hat{x}^n$'s s.t $\hat{x}^n y^n \in A_{n,\delta}$

Pf: similar to that in AEP.

Thm [Joint AEP]

Using above defs, $\forall \varepsilon > 0$ $\forall \delta > 0$ $\exists n_0$ s.t. $\forall n \geq n_0$

① $\Pr(A_{n,\delta}) \geq 1-\varepsilon$

② $(1-\varepsilon) 2^{n(H(XY)-\delta)} \leq |A_{n,\delta}| \leq 2^{n[H(XY)+\delta]}$

③ Suppose $x^n y^n$ is drawn according to the following distribution

$$q(x^n y^n) = p(x^n) \cdot p(y^n)$$

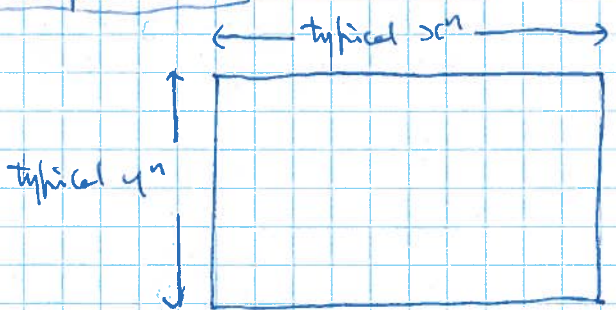
then $2^{-n(I(X:Y)+3\delta)} \leq \Pr_q (x^n y^n \in A_{n,\delta}) \leq 2^{-n(I(X:Y)-3\delta)}$

Pf (in Cover & Thomas Ch 8)

Ingredients:

The AEP itself, the union bound, etc.

Keep on board



Total $\approx 2^{nH(X)} \times 2^{nH(Y)}$ entries

The entry associated with $x^n y^n$ is

$$\begin{cases} 1 & \text{if } x^n y^n \in A_{n,\delta} \\ 0 & \text{otherwise.} \end{cases}$$

① says that tolerating a prob of failure ε, we can focus on this table

② says there are $\approx 2^{nH(XY)}$ "1"'s

Obs says each column has $\approx 2^{nH(Y|X)}$ "1"'s

each row . . . $2^{nH(X|Y)}$ "1"'s

③ says a <u>random</u> entry in the table has prob $= 2^{-nI(X:Y)}$ to be "1".

Example: $\Omega_X = \Omega_Y = \{0,1\}$

$p(00) = \frac{1}{2}(1-e)$

$p(01) = \frac{1}{2}e$

$p(10) = \frac{1}{2}e$

$p(11) = \frac{1}{2}(1-e)$

for $e \in [0,1]$, say $e = 0.1$.

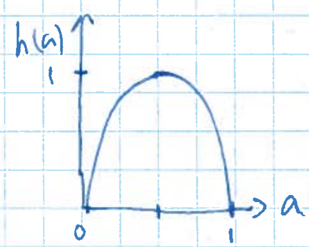
$H(X) = H(Y) = 1$

both marginals uniform

$H(XY) = 2\left(\frac{-1}{2}\right)(1-e)\log\left(\frac{1}{2}(1-e)\right) + 2\left(-\frac{e}{2}\right)\log\left(\frac{e}{2}\right)$

$= 1 + h(e) = 1.469$

where $h(a) = -a\log a - (1-a)\log(1-a)$ is the binary entropy function



$H(X|Y) = H(XY) - H(Y) = 0.469$.

$I(X:Y) = H(X) + H(Y) - H(XY) = 0.531$

$\therefore$ There are $\approx 2^{n\,1.469}$ jointly typical $x^n y^n$'s.

For each typical y^n, there are $\approx 2^{nH(X|Y)} = 2^{n\,0.469}$ $\tilde{x}^n$'s such that $\tilde{x}^n y^n$ is jointly typical.

For an $\tilde{x}^n$ chosen randomly for the typical set for $X_1 \ldots X_n$,

$\text{Prob}(\tilde{x}^n y^n \in A_{n,\delta}) \approx \frac{2^{nH(X|Y)}}{2^{nH(X)}} \doteq 2^{-nI(X:Y)} \approx 2^{-0.531n}$.

App 1: Distributed source coding

Goal: sample XY n times iid, get $x^n y^n$.

Give x^n to Alice, y^n to Bob.

How many cbits from Alice to Bob is needed for Bob to learn x^n?

Ans: $\approx n(H(X|Y) + \alpha)$ bits (and let's see what α should be).

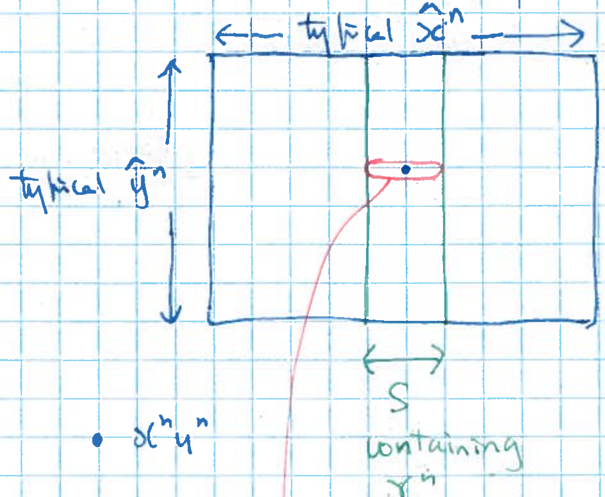
Method: Slepian-Wolf coding. Let T = Thed for $X_1 X_2 \ldots X_n$.

① Alice & Bob pre agree on a partition of T into $\approx 2^{n(H(X|Y)+\alpha)}$ sets each with at most $\underbrace{\frac{2^{n(H(X)+\delta)}}{2^{n(H(X|Y)+\alpha)}}}_{\boxed{\text{by AEP.}}} \approx 2^{n(I(X:Y)+\delta-\alpha)}$ elements.

② When Alice receives x^n, she tells Bob which of these $2^{n(H(X|Y)+\alpha)}$ sets contain x^n. Call the set S.

③ When Bob receives the label for S, he finds all $\hat{x}^n \in S$ that are jointly typical with y^n. Call this set J. Output "ERR" if $|J| = 0$ or $|J| > 1$.

Why this works? For large enough n, with prob $\geq 1-\varepsilon$, $x^n y^n \in A_{n,\varepsilon}$.



$\leftarrow$ $(\hat{x}^n, \hat{y}^n)$ entry $= \begin{cases} 1 & \text{if } \hat{x}^n \hat{y}^n \in A_{n,\varepsilon} \\ 0 & \text{otherwise} \end{cases}$

Total $\approx 2^{n(H(X)+H(Y))}$ entries

Only $\approx 2^{nH(XY)}$ 1's.

Bob checks all $\hat{x}^n y^n$ that are jointly typical.

The next step relies on how T is partitioned into the sets.

Idea: from all possible partition, pick one at random (say P) and analyse the prob P works.

In this case, S contains x^n, but all other elmts are chosen at random. Each such $\hat{x}^n$ has a prob

$$\approx \frac{2^{nH(X|Y)}}{2^{nH(X)}} \approx 2^{nI(X:Y)}$$ to be jointly typical with y^n.

1's in the row labeled by y^n # entries in the row.

More precisely, $\hat{x}^n$ & y^n are independent

$$\therefore \operatorname{Prob}_P(\hat{x}^n y^n \in A_{n,\delta}) \leq 2^{-n(I(X:Y)-3\delta)}$$ from the JAEP.

$$\therefore \operatorname{Prob}_P(\exists\, \hat{x}^n \in S,\ \hat{x}^n \neq x^n,\ \hat{x}^n y^n \in A_{n,\delta})$$

$$\underbrace{\leq}_{\text{union bound over each elt in } S} |S| \cdot 2^{-n(I(X:Y)-3\delta)} \leq 2^{n(4\delta-\alpha)}.$$

$\therefore$ Choosing $\alpha = 5\delta$, prob(P fails) $\leq \underbrace{\varepsilon}_{x^n y^n \notin A_{n,\delta}} + \underbrace{2^{-n\delta}}_{S \text{ bad}}$ (indep of y^n).

$\therefore \exists$ partition that works, requiring $n(H(X|Y)+5\delta)$ cbits.