Topic 3: Classical communication via classical channels

Let \( X \), \( Y \) be two RV's with joint distribution \( p(x,y) \) and sample space \( \Omega_x \times \Omega_y \).

1. \( H(XY) = - \sum_{x,y} p(x,y) \log p(x,y) \) (as defined earlier)

2. \( \forall y \in \Omega_y \), let \( p(y) = \sum_x p(x,y) \) the marginal distribution on \( Y \).

The conditional distribution of \( X \) given \( Y = y \):

\[
q_y(x) = \frac{p(x,y)}{p(y)} \quad \text{if} \ p(y) \neq 0
\]

**Def.** [Conditional entropy]

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

\[
H(X|Y) := \sum_y p(y) \left\{ H\left( q_y \right) \right\}
\]

- **Entropy of \( X \)**
  - **given \( Y = y \)**
  - **Average over \( Y \)**

**Thm.** (Chain rule)

\[
H(XY) = H(Y) + H(X|Y).
\]

**Pf.**

\[
H(XY) = - \sum_x \sum_y p(x,y) \log p(x,y)
\]

\[
= - \sum_y \left( \sum_x p(x,y) \log p(y) \right) - \sum_y \sum_x p(y) q_y(x) \log q_y(x)
\]

\[
= H(Y) + H(X|Y)
\]

(Or:

\[
H(X) = H(X|Y) + H(Y|X
\]

\[
H(X) = H(X|Y) + H(Y|X)
\]

\[
H(XY) = H(X) + H(Y|X)
\]
Def: [Relative entropy] [Kullback-Leibler distance]

Let \( p(x) \), \( q(x) \) be two distributions on \( \mathbb{R} \).

\[
D(p \| q) := \sum_{x \in \mathbb{R}} p(x) \log \frac{p(x)}{q(x)}
\]

NB: \( \sum p \), \( \sum q \) then \( D(p \| q) = \infty \), \( p \log \frac{p}{q} = \infty \).

NB: \( D \) non negative, \( D = 0 \iff p = q \), but not symmetric, no triangle inequality (not metric).

Not used in our course:

- \( D(p \| q) \) measures the inefficiency of assuming \( q \) when it's \( p \).
- In hypothesis testing, \( p^h \) is either \( p \) or \( q \) and iid samples are given. Any algorithm to discriminate \( p, q \), with \( m \) deterministic, and

\[
\text{output } \begin{cases} 0 \quad \text{if } x \in A & \text{for some } A \subseteq \mathbb{R}^n. \\ 1 \quad \text{otherwise} \end{cases}
\]

subject to the constraint \( \text{prob} (\text{out} = 0 \mid p) > 1 - \epsilon \).

let \( \beta(n, \epsilon) := \min \frac{1}{8} \log \left( \frac{1}{\epsilon} \right) \sim 2^{-n D(p \| q)} \),

\( p(A) > 1 - \epsilon \text{ under } p \).

Proof similar to AEP.

Def: [Mutual Info]

\[
I(X; Y) := D(p(x, y) \| p(x)p(y)) = - \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]
Properties of $H$, $D$, & $I$

1. Range: $0 \leq H(X) \leq \log |2\pi e|$

2. $H(X | Y) \geq 0$ (from $H_2$ & $H_3$, cumu combination of entropies)

3. $H(X | Y) > H(X)$ (from $H_2$ & chain rule). The more the merrier/messer

4. $D(p || f) \geq 0$.

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

Then: $-D(p || f) = -\sum_{x \in \text{supp}(p)} p(x) \log \frac{p(x)}{q(x)}$

$= \sum_{x \in \text{supp}(p)} p(x) \log \frac{q(x)}{p(x)}$

$= \sum_{x \in \text{supp}(p)} p(x) \log \frac{p(x)}{q(x)} \frac{q(x)}{p(x)}$

$= \sum_{x \in \text{supp}(p)} p(x) \log \sum_{x \in \text{supp}(p)} p(x) \frac{q(x)}{p(x)}$

$= 0$

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

$\text{iff } p(x) = p(0) \forall x$

5. $I(X; Y) \geq 0$, "=" iff $X, Y$ independent.
\[ I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \]

**Chain Rule**

**Interpretation:** \( I(X;Y) \) = decrease in ignorance of \( X \) given \( Y \) = amount of info of \( X \) carried by \( Y \).

**Note:** 6 follows from the def's of \( H \) & \( I \) but is NOT a definition.

**Pf:**
\[
I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]
\[
= \sum_{x,y} p(x,y) \log \frac{\frac{p(x,y)}{p(x)}}{\frac{p(y)}{p(x)}}
\]
\[
= \sum_{x} \left( \sum_{y} \frac{p(x,y)}{p(x)} \log \frac{p(x,y)}{p(y|x)} \right) + \sum_{y} \frac{p(y)}{p(x)} \log \frac{p(y)}{p(y|x)}
\]
\[
= H(X) - H(X|Y).
\]

**6. Subadditivity (SA)**

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

**Pf:** follows from 5 & 6.

**6.8 Conditioning reduces entropy**

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

**Pf:** from 5 & 48.
For 3 rv's $X,Y,Z$, $H(X|Z) > H(X|Y|Z)$.

**Pf:** Let $p(x|y,z)$ be the joint distn for XYZ.

For each $z$, let $q_z(x|y) = \frac{p(x,y,z)}{p(z)}$ where $p(z) = \sum_x p(x,y,z)$

Let $\hat{x}$ has distn $q_x(x|y)$.

We have $H(\hat{x}) \geq H(\hat{x}|y)$ from previous discussion.

- $H(\hat{x}) = \text{Entropy on } q_x(x) = \sum_x q_x(x) log_2(q_x(x))$

But $q_x(x) = \sum_y p(x,y,z) = \frac{p(x,z)}{p(z)} = p_x(x)$ from $\sum_x p(x,y,z) = 1$.

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

- $H(\hat{x}|\hat{y}) = \sum_{\hat{y}} (\sum_z q_z(x|y)) \cdot H(\hat{x}|\hat{y}, \hat{z}=\hat{y})$

$H(\hat{x}|\hat{y}, \hat{z}=\hat{y}) = \text{Entropy on } q_z(x|y) = \frac{\sum_x p(x,y,z)}{\sum_z p(x,y,z)} = \frac{p(x,y,z)}{p(x,z)} = \text{distn on } Y$

$\sum_x q_z(x|y) = \sum_x \frac{p(x,y,z)}{p(z)} = \sum_x \frac{p(x,y,z)}{p(x,z)} \cdot p(x,z) = \sum_x p(y|z) = \text{prob of } Y \text{ given } Z=z$.

$\therefore H(\hat{x}|\hat{y}) = H(X|Y,Z=z)$.
Now take convex combination \( \frac{\sum \psi(a) \cdot H(x)}{\sum \psi(a)} \geq \sum \frac{\psi(a)}{\sum \psi(a)} H(x|y)
\)

\begin{align*}
\frac{\psi(a)}{\sum \psi(a)} & H(x) & \geq & \frac{\psi(a)}{\sum \psi(a)} H(x|y) \\
H(x|y) & \geq & H(x|y|z) \quad & (\text{I10}) \\
\end{align*}

Def: The conditional mutual information

\[
I(x; y | z) = H(x | z) - H(x | y | z)
\]

\[
= H(y | z) - H(y | x | z)
\]

By (H8), \( I(x; y | z) \geq 0 \)
Concavity: mixing increases entropy

Let $\alpha_k > 0$, $\sum_{k=1}^{n} \alpha_k = 1$, $p_1, \ldots, p_n$ distributions on $\mathcal{X}$

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

Pf: Let $XX$ be rv's on $\{1, 2, \ldots, t\} \times \mathcal{X}$

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

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

$H(\sum_{k=1}^{n} \alpha_k p_k) = H(X)$

$\therefore \text{H10 follows from H8}$

Strong subadditivity SSA or data processing inequality DPI

- Def: We say $X \rightarrow Y \rightarrow Z$ (a Markov Chain)
  
  If $p(x, y, z) = p(x)p(y|x)p(z|y)$
  
  (i.e. If $p(z|y) = p(z|x, y)$)

- DPI: If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$

Pf: $I(X; YZ)$

$= H(X) + H(Z) - H(XZ) + H(XZ) - H(Z) - (H(XYZ) - H(YZ))$

$= I(X; Z) + H(XZ) - H(X, YZ) \geq I(X; Z)$ \hspace{1cm} (H10)

$I(X; YZ)$

$= H(X) + H(Y) - H(XY) + H(XY) - H(Y) - [H(XYZ) - H(YZ)]$

$= I(X; Y) + H(Z|Y) - H(Z|X) = I(X; Y)$ \hspace{1cm} (since $p(Z|Y) = p(Z|X)$)

$\therefore I(X; Y) \geq I(X; Z)$
In particular, if we process Y to obtain Z for any X that may be correlated with Y, we have $p(Z|Y) = p(Z|X,Y)$. So such processing can increase mutual info (from $I(X:Y)$ to $I(X:Z)$).
Def [Jointly typical sequence]
Given a distribution \( p(x,y) \), drawn iid \( n \) times
\( x^n y^n \) is jointly typical if

(a) \( -\frac{1}{n} \log p(x^n) - H(x) \leq d \) \( (x^n \text{ typical}) \)

(b) \( -\frac{1}{n} \log p(y^n) - H(y) \leq \delta \) \( (y^n \text{ typical}) \)

(c) \( -\frac{1}{n} \log p(x^n y^n) - H(x,y) \leq d \) \( (x^n y^n \text{ typical}) \)

Def [Jointly typical set]
\( \text{An} \cdot \text{id} = \{ x^n y^n \in \mathbb{R}_x^n \times \mathbb{R}_y^n : x^n y^n \text{ jointly typical} \} \)

Obs: If \( x^n y^n \in \text{An} \cdot \text{id} \),
then
\[
p(x^n | y^n) = \frac{p(x^n y^n)}{p(y^n)} \leq \frac{2^{-n(H(x,y)+\delta)}}{2^{-n(H(y)+\delta)}} = 2^{-n(H(x,y)+2\delta)}
\]

\[
p(x^n | y^n) = \frac{p(x^n y^n)}{p(y^n)} \geq \frac{2^{-n(H(x,y)+d)}}{2^{-n(H(y)+d)}} = 2^{-n(H(x,y)+2d)}
\]

NB. For this \( y^n \), there are \( 2^{-nH(x|y)} \) sets \( x^n \) s.t. \( x^n y^n \in \text{An} \cdot \text{id} \)

Pf: similar to that in AEP.
Thin [Joint AEP]

Using above defs, \( \forall \varepsilon > 0 \, \forall \delta > 0 \, \exists n \) st. \( \forall n \geq n_0 \).

1. \( \Pr (\text{Amid}) \geq 1 - \varepsilon \)
2. \( 2^{-n(H(X)+\delta)} \leq \Pr (\text{Amid}) \leq 2^{-n(H(X)+\delta)} \)
3. Suppose \( x^n y^n \) is drawn according to the following distribution \( f(x^n y^n) = 1^n 1^n \).

Then \( \sum 2^{-n(I(X,Y)+3\delta)} \leq \Pr (x^n y^n \in \text{Amid}) \leq \sum 2^{-n(I(X,Y)+3\delta)} \)

\( \Pr (\ln \text{lower} \& \text{Thomas 4.8}) \)

Ingredients:
- The AEP itself, the union bound, etc.

\( \sqrt{\text{Keep on board}} \)

\( \begin{array}{c}
\text{Typical x}^n \\
\text{Typical y}^n
\end{array} \)

Total \( 2^{nH(X)} \times 2^{nH(Y)} \) entries. The entry associated with \( x^n y^n \) is

\( \begin{cases} 1 & \text{if } x^n y^n \in \text{Amid} \\ 0 & \text{otherwise} \end{cases} \)

(a) Say that tolerating a prob \( 1 - \varepsilon \) for \( \varepsilon \), we can focus on this table

(b) Say there are \( 2^{nH(X)} \) \( 1 \)'s

(b) Say each column has \( 2^{nH(Y)} \) \( 1 \)'s

(c) Say each row has \( 2^{nH(Y)} \) \( 1 \)'s

(d) Say a random entry in the table has prob \( 2^{-nI(X;Y)} \) to be \( "1" \).
Example: \( \mathbb{R} = \mathbb{Z} = \{ 0, 1 \} \)

\[
P(00) = \frac{1}{2} (1-e)
\]

\[
P(01) = \frac{1}{2} e
\]

\[
P(10) = \frac{1}{2} e
\]

\[
P(11) = \frac{1}{2} (1-e)
\]

For \( e \in [0, 1] \), say \( e = 0.1 \).

\[
H(X) = \mu(H(Y)) = 1
\]

both marginals uniform

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

\[
= 1 + h(e) = 1.469
\]

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

\[
H(XY) = H(X) - H(Y) = 0.469
\]

\[
I(X; Y) = H(X) + H(Y) - H(XY) = 0.531
\]

\[
\text{There are } \approx 2^{-nH(XY)} \text{ joint typical } X^n Y^n \text{ 's.}
\]

For each typical \( Y^n \), there are \( \approx 2^{-nH(X|Y)} = 2^{-n0.469} \approx 2^{0.469} \text{ } \tilde{X}^n \text{ 's}
\]
such that \( \tilde{X}^n Y^n \) is jointly typical.

For an \( \tilde{X}^n \) chosen randomly from the typical set for \( X_1 \ldots X_n \),

\[
\text{Prob} \left( \tilde{X}^n Y^n \in A_{111} \right) \approx \frac{2^{-nH(X|Y)}}{2^{-nH(X)}} = 2^{-nI(X;X Y)} \approx 2^{-0.531n}
\]
**App 1:** Distributed source coding

**Goal:** Sample $XY$ $n$ times iid, get $X^nY^n$.
Give $X^n$ to Alice, $Y^n$ to Bob.

How many bits from Alice to Bob is needed for Bob to learn $X^n$?

**Ans:** $\approx n(H(XY) + 2)$ bits (and let's see what $d$ should be)

**Method:** Slepian-Wolf coding, let $T = \text{this for } X_1X_2...X_n$.

1. Alice & Bob agree on a partition of $T$ into $2^n(H(XY) + d)$ sets each with at most $\frac{2^n(H(XY) + d)}{2^n(H(XY) + 2)} \approx 2^{n(H(XY) + d - 2)}$ elements.

2. When Alice receives $X^n$, she tells Bob which of these $2^n(H(XY) + d)$ sets contain $X^n$. Call this set $S$.

3. When Bob receives the label for $S$, he finds all $Y^n \in S$ that are jointly typical with $X^n$. Call this set $\hat{S}$. 

```
\text{Output} \ "\text{ERR}\" \ \text{if} \ |\hat{S}| = 0 \ \text{or} \ |\hat{S}| > 1.
```

**Why this works?** For large enough $n$, with prob $\geq 1 - \epsilon$, $X^nY^n \approx \text{An.s.}$

Bob checks all $\hat{S} \subseteq Y^n$ 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 (seq \( p \)) and analyse the prob \( P \) works.

In this case, \( S \) contains \( x^n \), but all other events are chosen at random. Each such \( x^n \) has a prob
\[
\frac{2^n H(X|Y)}{2^n H(X)} \approx 2^{-n I(X:Y)}
\]
... to be jointly typical with \( y^n \).

More precisely, \( x^n \& y^n \) are independent
\[
P( x^n \& y^n \& \text{and}) \leq 2^{-n (H(X,Y) - 3d)} \quad \text{from the LEP.}
\]
\[
P( x^n \not\in S, x^n \neq x^n, y^n \& \text{and})
\]
\[
\leq \sum_{S \in S} 2^{-n (H(X,Y) - 3d)} \leq 2^{-n (4d - d)}
\]

Union bound over each event in \( S \)

Choosing \( d = 5d \), \( P( \text{reps}) \leq 2^{-n/6} \) (indep of \( y^n \)).

I find this technique works,

requiring \( n (H(X,Y) + 5d) \) chits.