Rami Pellumbi (rp534)
Hadi Alzayer (ha366)

1. **Monotonicity of Entropy for Stationary Processes.**
Let $\{X_i\}_{i=1}^{\infty}$ be a stationary sequence of random variables. For $n \in \mathbb{N}$ we denote $X^n := (X_1, ..., X_n)$. Prove that:

(a) For any $i, n \in \mathbb{N}$ with $1 \le i \le n$, we have $H\left(X_n \mid X^{n-1}\right) \le H\left(X_i \mid X^{i-1}\right)$.

Solution.
Fix integer index $i \in [1, n]$. Then

$$H\left(X_i \mid X_1\right) \ge H\left(X_i \mid X_1, X_2\right) \ge \ldots \ge H\left(X_i \mid X^{i-1}\right)$$

from the property conditioning cannot increase entropy. Letting $X_j^n := (X_j, X_{j+1}, ..., X_n)$, from stationarity we obtain

$$H\left(X_i \mid X^{i-1}\right) = H\left(X_n \mid X_{n-i}^{n-1}\right).$$

From here use the property that conditioning cannot increase entropy once more to obtain

$$\begin{aligned} H\left(X_n \mid X^{n-1}\right) &\le H\left(X_n \mid X_{n-i}^{n-1}\right) \\ &= H\left(X_i \mid X^{i-1}\right). \end{aligned}$$

■

(b) For any $n \in \mathbb{N}$, we have

$$\frac{H(X^n)}{n} \le \frac{H(X^{n-1})}{n-1}.$$

Solution. Begin by noting

$$\frac{H(X^n)}{n} = \frac{H\left(X_n \mid X^{n-1}\right) + \sum_{i=1}^{n-1} H\left(X_i \mid X^{i-1}\right)}{n}. \tag{1}$$

Now note that

$$\begin{aligned} H\left(X_n \mid X^{n-1}\right) &= \frac{\sum_{i=1}^{n-1} H\left(X_n \mid X^{n-1}\right)}{(n-1)} && (2) \\ &\le \frac{\sum_{i=1}^{n-1} H\left(X_i \mid X^{i-1}\right)}{n-1} && \text{(from part (a))} \\ &= \frac{H(X^{n-1})}{n-1} && (3) \end{aligned}$$

Plugging back into (1),

$$\begin{aligned}\frac{H(X^n)}{n} &\leq \frac{\frac{H(X^{n-1})}{n-1} + H(X^{n-1})}{n} \\ &= \frac{H(X^{n-1}) + (n-1)H(X^{n-1})}{n(n-1)} \\ &= \frac{H(X^{n-1})}{(n-1)}\end{aligned}$$

■

(c) For any $n \in \mathbb{N}$, we have

$$\frac{H(X^n)}{n} \geq H\left(X_n \mid X^{n-1}\right).$$

Solution.

$$\begin{aligned}H\left(X_n \mid X^{n-1}\right) &= \frac{\sum_{i=1}^{n} H\left(X_n \mid X^{n-1}\right)}{n} \\ &\leq \frac{\sum_{i=1}^{n} H\left(X_i \mid X^{i-1}\right)}{n} && \text{(from (a))} \\ &= \frac{H(X^n)}{n}\end{aligned}$$

■

2. **Entropy in Bytes.**
Let $P \in \mathcal{P}(\mathcal{X})$ and denote by p the associated PMF. The units of the entropy

$$H_a(P) = -\sum_{x \in \mathcal{X}} p(x) \log_a p(x)$$

are bits if the logarithm is to the base of $a = 2$ and bytes if the base is $a = 256$. Express $H_{256}(P)$ in terms of $H_2(X)$.

Solution.
Note that $2^{\log_2 x} = 256^{\log_{256} x}$. By taking $\log_{256}$ on both sides we get $\frac{1}{8}\log_2 x = \log_{256} x$.Thus

$$\begin{aligned}H_{256}(P) &= -\sum_{x \in \mathcal{X}} p(x) \log_{256} p(x) \\ &= -\frac{1}{8}\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \\ &= \frac{1}{8} H_2(P)\end{aligned}$$

■

3. **A Measure of Correlation.**
Let X_1 and X_2 be identically distributed but not necessarily independent. Assume that X_1 is not a constant, i.e, $H(X_1) > 0$. Define

$$\rho := 1 - \frac{H\left(X_2 \mid X_1\right)}{H(X_1)}$$

and show that

(a)
$$\rho = \frac{I(X_1; X_2)}{H(X_1)}.$$

Solution.
First, note that $H(X_1) = H(X_2)$ since they are identically distributed.

$$\begin{aligned}
\rho &= 1 - \frac{H\left(X_2 \mid X_1\right)}{H(X_1)} \\
&= \frac{H(X_1) - H\left(X_2 \mid X_1\right)}{H(X_1)} \\
&= \frac{H(X_2) - H\left(X_2 \mid X_1\right)}{H(X_1)} && \text{(since } H(X_1) = H(X_2)\text{)} \\
&= \frac{I(X_1; X_2)}{H(X_1)}
\end{aligned}$$

■

(b) $0 \leq \rho \leq 1$

Solution.
Here, it suffices to show that

$$0 \leq H\left(X_2 \mid X_1\right) \leq H(X_1).$$

The non-negativity follows from properties of mutual information.

$$H\left(X_2 \mid X_1\right) \leq H(X_2) = H(X_1)$$

is obtained from the property that conditioning cannot increase entropy. ■

(c) Find a necessary and sufficient condition for $\rho = 0$.

Solution.
If $X_1 \perp\!\!\!\perp X_2$, then

$$H\left(X_2 \mid X_1\right) = H(X_2) = H(X_1),$$

giving

$$\frac{H\left(X_2 \mid X_1\right)}{H(X_1)} = 1$$

and thus $\rho = 0$. ■

(d) Find a sufficient condition for $\rho = 1$.

Solution.
Here, we need

$$H\left(X_1 \mid X_2\right) = 0.$$

This occurs if X_1 and X_2 are completely dependent on each other, which gives us $\rho = 1$ ■

4. **Random Questions.**
One wishes to learn the value of a random variable $X \sim P_X \in \mathcal{P}(\mathcal{X})$. A question $Q \sim P_Q \in \mathcal{P}(\mathcal{Q})$ is asked at random according to P_Q. This results in a answer $A := a(X, Q)$, where $a : \mathcal{X} \times \mathcal{Q} \to \mathcal{A}$ is a deterministic answer function that attaches an answer $a(x, q)$ to any value-question pair $(x, q) \in \mathcal{X} \times \mathcal{Q}$. Suppose that X and the question Q are independent (modeling the fact that the inquirer has no prior knowledge about X when asking Q). With respect to this model, $I(X; Q, A)$ is the information the question-answer pair (Q, A) conveys about X.

(a) Show that $I(X; Q, A) = H\left(A \mid Q\right)$ and interpret this result.

Solution.
We have

$$I(X; Q, A) = H(Q, A) - H(Q, A|X).$$

Expanding the expressions using the chain rule yields:

$$\begin{aligned} I(X; Q, A) &= H(Q, A) - H(Q, A|X) \\ &= H(Q) + H(A|Q) - H(Q|X) - H(A|Q, X) \end{aligned}$$

From $Q \perp\!\!\!\perp X$, we obtain $H(Q|X) = H(Q)$, and we use the fact that A is completely determined by Q, A, using the deterministic function a, to obtain $H(A|X, A) = 0$. The above thus simplifies to

$$\begin{aligned} I(X; Q, A) &= H(Q) + H(A|Q) - H(Q|X) - H(A|Q, X) \\ &= H(A|Q) \end{aligned}$$

■

(b) Now suppose that two i.i.d questions $Q_1, Q_2 \sim P_Q$ are asked, eliciting answers

$$A_1 := A(X, Q_1)$$

and

$$A_2 := A(X, Q_2).$$

Show that the two questions are less valuable than twice the value of a single question in the sense that

$$I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1).$$

Solution.
From the previous part, we have can express the RHS as $2H(A|Q)$. Expanding the LHS,

$$\begin{aligned}
I(X; Q_1, A_1, Q_2, A_2) &= H(Q_1, A_1, Q_2, A_2) - H(Q_1, A_1, Q_2, A_2|X) \\
&= H(Q_1) + H(A_1|Q_1) + H(Q_2|A_1, Q_1) + H(A_2|Q_2, A_1, Q_1) \\
&- H(Q_1|X) - H(A_1|Q_1, X) - H(Q_2|A_1, Q_1, X) - H(A_2|Q_2, A_1, Q_1, X) \\
&= H(A_1|Q_1) + H(A_2|Q_2, A_1, Q_1)
\end{aligned}$$

Since conditioning cannot increase entropy and the questions are i.i.d, we thus have

$$H(A_2|Q_2, A_1, Q_1) \leq H(A_1|Q_1),$$

giving us the desired result

$$H(A_1|Q_1) + H(A_2|Q_2, A_1, Q_1) \leq 2H(A_1|Q_1).$$

■

5. **Joint Letter Typical Set.**
Let $P_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a distribution with $|\mathrm{supp}(P_{X,Y})| < \infty$ and denote by $p_{X,Y}$ its PMF. For $n \in \mathbb{N}$ and $\epsilon > 0$ recall the definition of the joint-letter typical set

$$\mathcal{T}_\epsilon^{(n)}(P_{X,Y}) := \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |\nu_{x^n,y^n}(a,b) - p_{X,Y}(a,b)| < \epsilon p_{X,Y}(a,b), \quad \forall (a,b) \in \mathcal{X} \times \mathcal{Y}\right\}$$

where

$$\nu_{x^n,y^n}(a,b) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{(x_i,y_i)=(a,b)\}},$$

for $(a,b) \in \mathcal{X} \times \mathcal{Y}$, is the empirical frequency of the pair $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$. Prove the following properties:

(a) If $(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(P_{X,Y})$ then $x^n \in \mathcal{T}_\epsilon^{(n)}(P_X)$ and $y^n \in \mathcal{T}_\epsilon^{(n)}(P_Y)$.

Solution.

$$(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(P_{X,Y}) \implies |\nu_{x^n,y^n}(a,b) - p_{X,Y}(a,b)| < \epsilon p_{X,Y}(a,b), \quad \forall (a,b) \in \mathcal{X} \times \mathcal{Y}.$$

Expanding the absolute value yields

$$-\epsilon p_{X,Y}(a,b) < \nu_{x^n,y^n}(a,b) - p_{X,Y}(a,b) < \epsilon p_{X,Y}(a,b), \quad \forall (a,b) \in \mathcal{X} \times \mathcal{Y}$$

Summing over all $b \in \mathcal{Y}$, we obtain

$$\begin{aligned} &-\epsilon p_X(a) < \nu_{x^n}(a) - p_X(a) < \epsilon p_X(a), \quad \forall a \in \mathcal{X} \\ &\implies |\nu_{x^n}(a) - p_X(a)| < \epsilon p_X(a), \quad \forall a \in \mathcal{X} \\ &\implies x^n \in \mathcal{T}_\epsilon^{(n)}(P_X). \end{aligned}$$

Similarly we instead sum over all $a \in \mathcal{X}$, to obtain

$$\begin{aligned} &-\epsilon p_Y(b) < \nu_{y^n}(b) - p_Y(b) < \epsilon p_Y(b), \quad \forall b \in \mathcal{Y} \\ &\implies |\nu_{y^n}(b) - p_Y(b)| < \epsilon p_Y(b), \quad \forall b \in \mathcal{Y} \\ &\implies y^n \in \mathcal{T}_\epsilon^{(n)}(P_Y). \end{aligned}$$

■

(b) For any $(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(P_{X,Y})$, we have

i. $2^{-n(1+\epsilon)H(P_{X,Y})} \le P_{X,Y}^{\otimes n}(\{(x^n, y^n)\} \le 2^{-n(1-\epsilon)H(P_{X,Y})}$.

Solution.

This an immediate consequence of the Typical averaging lemma. Let

$$g(x,y) = -\log\left(p(x,y)\right).$$

Note that

$$\mathbb{E}_P[g(X,Y)] = H(P_{X,Y}) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i) = \frac{1}{n}\log\left(\frac{1}{P_{X,Y}^{\otimes n}(\{x^n, y^n\})}\right).$$

The lemma then implies

$$(1-\epsilon)H(P_{X,Y}) \le \frac{1}{n}\log\left(\frac{1}{P_{X,Y}^{\otimes n}(\{x^n, y^n\})}\right) \le (1+\epsilon)H(P_{X,Y}).$$

By taking the exponential of all terms, we get

$$2^{-n(1-\epsilon)H(P_{X,Y})} \ge P_{X,Y}^{\otimes n}(\{x^n\}) \ge 2^{-n(1+\epsilon)H(P_{X,Y})}.$$

■

ii. $2^{-n(1+\epsilon)H(P_X)} \le P_X^{\otimes n}(\{x^n\} \le 2^{-n(1-\epsilon)H(P_X)}$.

Solution.
This an immediate consequence of the Typical averaging lemma. Let $g(x) = -\log\big(p(x)\big)$. Note that $\mathbb{E}_P[g(X)] = H(P_X)$, and $\frac{1}{n}\sum_{i=1}^n g(x_i) = \frac{1}{n}\log\left(\frac{1}{P_X^{\otimes n}(\{x^n\})}\right)$. The lemma then implies

$$(1-\epsilon)H(P_X) \leq \frac{1}{n}\log\left(\frac{1}{P_X^{\otimes n}(\{x^n\})}\right) \leq (1+\epsilon)H(P_X).$$

By taking the exponential of all terms, we get

$$2^{-n(1-\epsilon)H(P_X)} \geq P_X^{\otimes n}(\{x^n\}) \geq 2^{-n(1+\epsilon)H(P_X)}.$$

■

iii. $2^{-n(1+\epsilon)H(P_Y)} \leq P_Y^{\otimes n}(\{y^n\} \leq 2^{-n(1-\epsilon)H(P_Y)}$.
Solution.
See (ii) but define $g(y)$ instead. ■

(c) If $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d according to $P_{X,Y}$, then

$$\lim_{n\to\infty} P_{X,Y}^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})) = 1.$$

Solution.
By definition we have

$$P_{X,Y}^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})\right) = P_{X,Y}^{\otimes n}\left(\bigcap_{(a,b)\in\mathcal{X}\times\mathcal{Y}} \big\{(x^n, y^n) : |\nu_{x^n,y^n}(a,b) - p_{X,Y}(a,b)| \leq \epsilon p_{X,Y}(a,b)\big\}\right).$$

By the weak law of large number, we have that for a set of arbitrary functions $\{f_k(x,y)\}_{k=1}^\infty$, (each with finite expectation) and any $\delta > 0$,

$$\lim_{n\to\infty} P^{\otimes n}\left(\bigcap_{k=1}^K \left\{(x,y) : \left|\frac{1}{n}\sum_{i=1}^n f_k(x_i,y_i) - \mathbb{E}_P\big[f_k(X,Y)\big]\right| \leq \delta\right\}\right) = 1.$$

Now, set $K = |\mathcal{X}| = |\mathcal{Y}|$, and $f_k = \mathbb{1}_{\{(a,b)\}}$, for each $(a,b) \in \mathcal{X}\times\mathcal{Y}$. Then,

$$\mathbb{E}_P[f_k(X,Y)] = p_{X,Y}(a,b),$$

and

$$\frac{1}{n}\sum_{i=1}^n f_k(x_i,y_i) = \nu_{x^n,y^n}(a,b).$$

The WLLN then implies

$$\lim_{n\to\infty} P_{X,Y}^{\otimes n}\left(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})\right) = \lim_{n\to\infty} P_{X,Y}^{\otimes n}\left(\bigcap_{(a,b)\in\mathcal{X}\times\mathcal{Y}} \left\{(x^n,y^n) : |\nu_{x^n,y^n}(a,b) - p_{X,Y}(a,b)| \le \epsilon p_{X,Y}(a,b)\right\}\right)$$
$$= 1.$$

■

(d) The cardinality of $\mathcal{T}_\epsilon^{(n)}(P_{X,Y})$ is bounded as

$$(1-\delta)2^{n(1-\epsilon)H(P_{X,Y})} \le |\mathcal{T}_\epsilon^{(n)}(P_{X,Y})| \le 2^{n(1+\epsilon)H(P_{X,Y})}$$

where the lower bound holds for any $\delta > 0$ and n large enough.

Solution.
Use the fact that $\mathcal{T}_\epsilon^{(n)}(P_{X,Y}) \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ to obtain

$$\begin{aligned}1 = P_{X,Y}^{\otimes n}(\mathcal{X}^n\times\mathcal{Y}^n) \ge P_{X,Y}^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})) &= \sum_{(x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(P_{X,Y})} P_{X,Y}^{\otimes n}(\{x^n,y^n\}) \\ &\ge |\mathcal{T}_\epsilon^{(n)}(P_{X,Y})| \cdot 2^{-n(1+\epsilon)H(P_{X,Y})} \\ &\implies |\mathcal{T}_\epsilon^{(n)}(P_{X,Y})| \le 2^{n(1+\epsilon)H(P_{X,Y})}.\end{aligned}$$

To obtain the lower bound, let n be sufficiently large so that $P_{X,Y}(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})) = 1$. Then for any $\delta > 0$,

$$1-\delta \le P_{X,Y}^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})).$$

Thus

$$\begin{aligned}1-\delta &\le \sum_{(x^n,y^n)\in\mathcal{T}_\epsilon^{(n)}(P_{X,Y})} P_{X,Y}^{\otimes n}(\{x^n,y^n\}) \le |\mathcal{T}_\epsilon^{(n)}(P_{X,Y})| \cdot 2^{-n(1-\epsilon)H(P_{X,Y})} \\ &\implies (1-\delta)2^{n(1-\epsilon)H(P_{X,Y})} \le |\mathcal{T}_\epsilon^{(n)}(P_{X,Y})|.\end{aligned}$$

■

6. **Mismatch Letter-Typicality.**
Let $n \in \mathbb{N}, \epsilon > 0, X^n \sim P^{\otimes n}$, and $Q \ll P$.

(a) Prove that

$$(1-\epsilon)2^{-n(D_{\mathsf{KL}}(Q\|P)+\delta(\epsilon))} \le P^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q)) \le 2^{-n(D_{\mathsf{KL}}(Q\|P)-\delta(\epsilon))}$$

where $\lim_{\epsilon\to 0}\delta(\epsilon) = 0$ and the lower bound holds for any n large enough. Provide an explicit expression for $\delta(\epsilon)$.

Solution.
Here, we use the typical averaging lemma with respect to typical letter sequence from Q:

$$(1-\epsilon)\mathbb{E}_Q[g(X)] \leq \frac{1}{n}\sum_{i=1}^{n} g\left(x_i\right) \leq (1+\epsilon)\mathbb{E}_Q[g(X)].$$

Letting $g(x) = \log\frac{q(x)}{p(x)}$, so we have $\mathbb{E}_Q[g(X)] = D_{\mathsf{KL}}\left(Q\|P\right)$, allows us to rewrite the above as

$$n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right) \leq \sum_{i=1}^{n}\log\frac{q(x_i)}{p(x_i)} \leq n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right). \qquad (\forall x^n \in \mathcal{T}_\epsilon^{(n)}(Q))$$

Since the probability of each element in x^n is sampled independently,

$$\log P^{\otimes n}(x^n) = \sum_{i=1}^{n}\log(P(x_i)),$$

giving

$$n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right) \leq \log\frac{Q^{\otimes n}(x^n)}{P^{\otimes n}(x^n)} \leq n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right).$$

By multiplying by -1 and exponentiation all sides, we obtain

$$Q^{\otimes n}(x^n)2^{-n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)} \leq P^{\otimes n}(x^n) \leq Q^{\otimes n}(x^n)2^{-n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)}.$$

To find $P^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q))$, then we sum over all $x^n \in \mathcal{T}_\epsilon^{(n)}(Q)$, obtaining

$$Q^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q))2^{-n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)} \leq P^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q)) \leq Q^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q))2^{-n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)}.$$

For the upper bound, we have

$$Q^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q))2^{-n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)} \leq 2^{-n(1-\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)}.$$

For the lower bound, we use the property that

$$Q^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q) \xrightarrow[n\to\infty]{} 1.$$

Thus, for large enough n, we have

$$Q^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q))2^{-n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)} \geq (1-\epsilon)2^{-n(1+\epsilon)D_{\mathsf{KL}}\left(Q\|P\right)}.$$

To conclude,

$$(1-\epsilon)2^{-n(D_{\mathsf{KL}}\left(Q\|P\right)+\delta(\epsilon))} \leq P^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(Q)) \leq 2^{-n(D_{\mathsf{KL}}\left(Q\|P\right)-\delta(\epsilon))},$$

where we have $\delta(\epsilon) = D_{\mathsf{KL}}\left(Q\|P\right)\epsilon$ which indeed goes to 0 as $\epsilon \to 0$. ■

(b) Deduce that for $P_{X,Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with marginals P_X and P_Y, we have

$$(1-\epsilon)2^{-nI(X;Y)+\tilde{\delta}(\epsilon))} \leq P_X^{\otimes n} \otimes P_Y^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P_{X,Y})) \leq 2^{-n(I(X;Y)-\tilde{\delta}(\epsilon))}.$$

What is $\tilde{\delta}(\epsilon)$ in this case?

Solution.
This follows immediately from the previous part, since mutual information by definition can be written as

$$I(X;Y) = D_{\mathsf{KL}}\left(P_{XY} \| P_X \otimes P_Y\right).$$

This gives us

$$(1-\epsilon)2^{-n(1+\epsilon)I(X;Y)} \leq P_X^{\otimes n} \otimes P_Y^{\otimes n}(\mathcal{T}_\epsilon^{(n)}(P_{XY})) \leq 2^{-n(1-\epsilon)I(X;Y)}.$$

Similarly to previous part, we get $\hat{\delta}(\epsilon) = I(X;Y)\epsilon$ which approaches zero as $\epsilon \to 0$. ■

7. **Discrete Memoryless Channel Without Feedback.**
Consider the communication over a noisy channel scenario as described by the induced distribution on $\mathcal{M} \times \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{M}$:

$$P_{M,X^n,Y^n,\hat{M}}(m, x^n, y^n, \hat{m}) = P_M(m)\mathbb{1}_{x^n=f_n(m)} P_{Y|X}^{\otimes n}(y^n \mid x^n)\mathbb{1}_{\hat{m}=g_n(y^n)},$$

where $P_M \in \mathcal{P}(\mathcal{M})$ is a message distribution and $c_n := (f_n, g_n)$ is a code (encoder-decoder pair). Assume that:

(i) The channel is memoryless, i.e., there exists a (single-letter) transition kernel $P_{Y|X}$ such that

$$P^{(c_n)}(y_i \mid m, x^i, y^{i-1}) = P_{Y|X}(y_i \mid x_i),$$

for all $i = 1, ..., n$.

(ii) The channel is without feedback, i.e.,

$$P^{(c_n)}(x_i \mid m, x^{i-1}, y^{i-1}) = P^{c_n}(x_i \mid m, x^{i-1}),$$

for all $i = 1, ..., n$.

Prove that $P^{(c_n)}(y^n \mid x^n) = \prod_{i=1}^{n} P(Y \mid X)(y_i \mid x_i)$.

Solution.
We have

$$P^{(c_n)}(y^n \mid x^n) = \sum_m P(m)P(y^n|x^n, m),$$

but since the probability is non-zero iff $x^n = f(m)$, we need to make sure that we are not conditioning on zero probability events. To do this we rewrite the sum as

$$\sum_m P(m|x^n = f(m))P(y^n|x^n, m).$$

Note that because of the no-feedback property, and the deterministic encoding function f, we conclude x^n to be a deterministic function of m. Thus

$$P(y^n|x^n, m) = P(y^n|m).$$

Using the chain rule, we have

$$\begin{aligned}
\sum_m P(m|x^n = f(m))P(y^n|x^n, m) &= \sum_m P(m|x^n = f(m)) \prod_{i=1}^{n} (y_i|y^{i-1}, x^n, m) && \text{chain rule} \\
&= \sum_m P(m|x^n = f(m)) \prod_{i=1}^{n} (y_i|y^{i-1}, x^i, m) && x^n \text{ determined by } m \\
&= \sum_m P(m|x^n = f(m)) \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) && \text{memoryless property} \\
&= \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) && \text{marginalizing over } m
\end{aligned}$$

■

8. **Capacity of Binary Erasure Channel.**
Consider the binary erasure channel (BEC) in which a fraction $\alpha \in [0, 1]$ of the transmitted bits are lost (erased) as depicted in Figure 1. More precisely, the BEC of parameter α is specified by the tuple $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, where $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, e\}$ and $P_{Y|X}$ is described by the relation:

$$Y = \begin{cases} X, & \text{w.p.} 1 - \alpha, \\ e, & \text{w.p. } \alpha. \end{cases}$$

Find a closed form expression that depends only on α for the capacity $\max_{P_X} I(X; Y)$ of this BEC.
Hint: Consider the function $E = \mathbb{1}_{Y=e}$ and show that $I(X; Y) = I(X; Y, E) = I(X; Y \mid E)$.

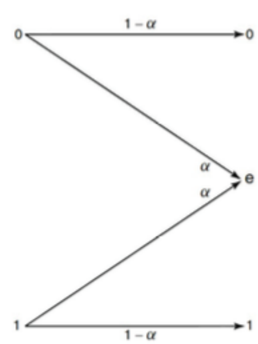

Fig. 1: Binary erasure channel.

Solution.
By definition

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

Given X, we conclude Y to essentially be a Bernoulli measure, where

$$\mathbb{P}(Y=?|X=x) = \alpha \quad \text{and} \quad \mathbb{P}(Y=x|X=x) = (1-\alpha).$$

Thus $H(Y|X) = H_B(\alpha)$ and doesn't depend on X. To maximize mutual information, we need to maximize $H(Y)$. First we expand the entropy definition by finding the probability of different events for Y.

$$\begin{aligned}\mathbb{P}(Y=?) &= \alpha \\ \mathbb{P}(Y=0) &= (1-\alpha)\mathbb{P}(X=0) \\ \mathbb{P}(Y=1) &= (1-\alpha)\mathbb{P}(X=1)\end{aligned}$$

By plugging these probabilities into the definition of entropy we get

$$\begin{aligned}H(Y) &= \alpha\log\frac{1}{\alpha} + (1-\alpha)\mathbb{P}(X=0)\log\frac{1}{(1-\alpha)\mathbb{P}(X=0)} + (1-\alpha)\mathbb{P}(X=1)\log\frac{1}{(1-\alpha)\mathbb{P}(X=1)} \\ &= \alpha\log\frac{1}{\alpha} + (1-\alpha)(\log\frac{1}{1-\alpha} + (1-\alpha)\mathbb{P}(X=0)\log\frac{1}{\mathbb{P}(X=0)} + \mathbb{P}(X=1)\log\frac{1}{\mathbb{P}(X=1)}) \\ &= H_B(\alpha) + (1-\alpha)H(X)\end{aligned}$$

Plugging back into the definition for mutual information gives us the expression

$$\begin{aligned}I(Y;X) &= H(Y) - H(Y|X) \\ &= H_B(\alpha) + (1-\alpha)H(X) - H_B(\alpha) \\ &= (1-\alpha)H(X).\end{aligned}$$

To maximize mutual information, we need to maximize $H(X)$. Since X is a Bernoulli measure, we maximize it by setting $X = Ber(\frac{1}{2})$ giving us $H(X) = 1$.
In conclusion, we get

$$C_{\max} = \max P_X I(Y;X) = (1-\alpha)$$

■

9. **Capacity of Noisy Typewriter.**
Suppose we have a malfunctioning typewriter that we model as a channel from the keystroke X_{in} to the typed symbol Y_{out}. Specifically, let $\mathcal{X}_{in} = \mathcal{Y}_{out} = \{A, B, C, ..., Z\}$ and define

$$\text{con} : \mathcal{X}_{in} \to \mathcal{Y}_{out}$$

as the function that (circularly) maps any letter of the alphabet to the next one, e.g., con(A) = B and con(Z)=A. The noisy typewriter channel is described by the relation

$$Y_{out} = \begin{cases} X_{in}, & \text{w.p.}\frac{1}{2}, \\ \text{con}(X_{in}), & \text{w.p.}\frac{1}{2}. \end{cases}$$

In words, the keystroke X_{in} is either typed unaltered with probability $\frac{1}{2}$ or is transformed to the next letter of the alphabet with probability $\frac{1}{2}$. Find the capacity $\max_{P_X} I(X;Y)$ of the noisy typewriter channel.

Solution. We have

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

Note that given $X = x$, Y becomes a Bernoulli variable with parameter $p = \frac{1}{2}$ where

$$\mathbb{P}(Y = x|X = x) = \frac{1}{2}, \mathbb{P}(Y = con(x)|X = x) = \frac{1}{2},$$

giving $H(Y|X) = 1$. We now have

$$I(X;Y) = H(Y) - 1.$$

To maximize the mutual information, we need to pick a distribution P_X that maximizes $H(Y)$. Note that $H(Y)$ is maximum when Y is uniform. However, if we let $P_X = Unif(\mathcal{X})$, then we get Y to be a uniform distribution. Denote $prev(x)$ to be the inverse of $con(x)$ and $C = 1$ to be the event that $Y = con(X)$ and $C = 0$ otherwise. We get for each letter l,

$$\mathbb{P}(Y = l) = \mathbb{P}(X = l, C = 1) + \mathbb{P}(X = prev(l)|C = 0) = \frac{1}{2}\frac{1}{|\mathcal{X}|} + \frac{1}{2}\frac{1}{|\mathcal{X}|} = \frac{1}{|\mathcal{X}|}.$$

So we have $C_{max} = \log|\mathcal{X}| - 1$ ■