Indian Institute of Technology, Guwahati

Pre-Final Year : 6^{th} Semester

Information Theory & Coding

Prof. Tony Jacob

Written by
Animesh Renanse

February 13, 2021

Contents

1 Entropy, Relative Entropy & Mutual Information

1.1 Entropy

Entropy is the measure of *uncertainty* of a random variable.

★ **Definition 1.** (**Entropy**) The entropy $H(X)$ of a discrete random variable X is defined by

$$\boxed{H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)} \tag{1}$$

where $\mathcal{X}$ is the support of the variable X.

Remark. Few things to note:

- If the base of the logarithm is b, we denote the entropy as $H_b(X)$, if not, then it is assumed to be 2.
- If the base of the logarithm is e, then the entropy is measured in ***nats***.
- We also denote the following:

$$H_2(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

Now, discussion happens on the case when $g(X) = \log \frac{1}{p(X)}$ and the expected value of $g(X)$, which is $\mathbb{E}g(X)$.

Remark. The entropy of X can also be interpreted as the expected value of the random variable $\log \frac{1}{p(X)}$, where X is drawn according to pmf $p(x)$

$$H(X) = \mathbb{E}_p \log \frac{1}{p(X)}$$

Lemma 1. $H(X) \geq 0$.

Proof. Since $0 \leq p(X) \leq 1$, then $\log \frac{1}{p(X)} \geq 0$, hence $H(X)$ is always ≥ 0. ■

Lemma 2. $H_b(X) = (\log_b a) H_a(X)$.

Proof. Remember that $\log_b a = \frac{\log_b p}{\log_a p}$. ■

1.2 Joint Entropy & Conditional Entropy

Extending the Definition 1 to a *pair of random variables*.

★ **Definition 2.** (**Joint Entropy**) Let joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$\boxed{\begin{aligned} H(X,Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \\ &= -\mathbb{E}\left[\log p(X,Y)\right] \end{aligned}} \tag{2}$$

★ **Definition 3.** (**Conditional Entropy**) If $(X,Y) \sim p(x,y)$, then the conditional entropy $H(Y|X)$ is defined as

$$\boxed{\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x) \quad (\star) \\ &= -\mathbb{E}_{p(x,y)}\left[\log p(Y|X)\right] \end{aligned}} \tag{3}$$

Remark. This is just the expected value of the entropies of the conditional distribution.

⊛ **Theorem 1.** ***(Chain Rule)*** For a random vector (X, Y), we have

$$\boxed{H(X,Y) = H(X) + H(Y|X)} \tag{4}$$

Proof. It's trivial to see following by the basic properties of conditional distributions:

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

Hence proved. ■

→ **Corollary 1.** We also have that

$$\boxed{H(X,Y|Z) = H(X|Z) + H(Y|X,Z)}$$

Proof. Simply note :

$$\begin{aligned}
H(X,Y|Z) &= H(X,Y,Z) - H(Z) \qquad \text{(Theorem 1)} \\
&= H(Y|X,Z) + H(X,Z) - H(Z) \\
&= H(Y|X,Z) + H(X|Z)
\end{aligned}$$

Hence proved. ■

1.3 Relative Entropy & Mutual Information

★ **Definition 4.** (**Relative Entropy**) The relative entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\boxed{\begin{aligned} D\left(p\|q\right) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\ &= \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] \end{aligned}} \tag{5}$$

Remark. Few things to note:

- The relative entropy $D\left(p\|q\right)$ is the measure of the *inefficiency* of assuming that the distribution is q when the true distribution is p.
- As much it might be tempting to say KL-divergence as a distance, it's not, because it is not symmetric and doesn't follows triangle inequality. However, $D\left(p\|q\right) = 0 \iff p = q$.

★ **Definition 5.** (**Mutual Information**) Consider two random variables X and Y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between joint distribution $p(x,y)$ and

the product distribution $p(x)p(y)$. That is,

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D\left(p(x,y)\|p(x)p(y)\right) \\ &= \mathbb{E}_{p(x,y)}\left[\log\frac{p(x,y)}{p(x)p(y)}\right] \end{aligned} \tag{6}$$

Remark. Few things to note:

- This is a measure of amount of information one variable contains about the other. We can see this as we are measuring the *distance* between the assumption whether X and Y are independent to the true joint distribution.
- Hence it is also the reduction in uncertainty of one random variable due to the knowledge of the other, as shown by Theorem 2, 1^{st} eq.

$\boxed{\star}$ **Definition 6.** (**Conditional Mutual Information**) The conditional mutual information of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \mathbb{E}_{p(x,y,z)}\left[\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right] \end{aligned} \tag{7}$$

Remark. Conditional Mutual Information follows chain rule for Information (Theorem 4).

$\boxed{\star}$ **Definition 7.** (**Conditional Relative Entropy**) For joint probability mass function $p(x,y)$ and $q(x,y)$, the conditional relative entropy $D\left(p(y|x)\|q(y|x)\right)$ is the average of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. That is,

$$\begin{aligned} D\left(p(y|x)\|q(y|x)\right) &= \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{q(y|x)} \\ &= \mathbb{E}_{p(x,y)}\left[\log\frac{p(y|x)}{q(y|x)}\right] \end{aligned} \tag{8}$$

Remark. Note that the notation for conditional relative entropy $D\left(p(y|x)\|q(y|x)\right)$ does not include the fact that expectation is taken over X. So it must be assumed accordingly from the context.

1.4 Relationship between Entropy & Mutual Information

$\boxed{\circledast}$ **Theorem 2.** (***Mutual Information & Entropy***) The following are the relations between the two notions of information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ I(X;Y) &= H(Y) - H(Y|X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(X;Y) &= I(Y;X) \\ I(X;X) &= H(X) \end{aligned} \tag{9}$$

Proof. We have the following:

$$\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x|y)}{p(x)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) \\
&= -H(X|Y) + \sum_{x\in\mathcal{X}} p(x)\log p(x) \\
&= -H(X|Y) + H(X) \\
&= -H(X,Y) + H(Y) + H(X) \qquad \text{(Theorem 1)}
\end{aligned}$$

By symmetry, $I(X;Y) = I(Y;X)$ and since

$$\begin{aligned}
H(X|X) &= -\sum_{x\in\mathcal{X}} p(x)\log p(x|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log 1 \\
&= 0
\end{aligned}$$

therefore, $I(X;X) = H(X)$. ■

1.5 Chain Rules

Entropy of collection of random variables is the sum of conditional entropies!

Theorem 3. (***Chain Rule for Entropy***) Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then,

$$\boxed{H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)} \tag{10}$$

Proof. This is an easy consequence of Theorem 1 over collection of random variables:

$$\begin{aligned}
H(X_1, X_2) &= H(X_1) + H(X_2|X_1) \\
H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\
H(X_1, X_2, X_3) &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \qquad \text{(Corollary 1)} \\
&\vdots = \vdots \\
H(X_1, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1)
\end{aligned}$$

Hence proved. ■

Theorem 4. (**Chain Rule for Information**) We have the following for mutual information:

$$\boxed{I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, \ldots, X_1)} \tag{11}$$

Proof. This is a consequence of Theorem 2, 1^{st} equation:

$$\begin{aligned}
I(X_1, \ldots, X_n; Y) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1, Y) \qquad \text{(Theorem 3)} \\
&= \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1)
\end{aligned}$$

Hence proved. ■

⊛ **Theorem 5.** (**Chain Rule for Relative Entropy**) The relative entropy between two joint distributions on a pair of random variables can be expanded as the sum of a relative entropy and a conditional relative entropy. That is,

$$\boxed{D\left(p(x,y)\|q(x,y)\right) = D\left(p(x)\|q(x)\right) + D\left(p(y|x)\|q(y|x)\right)} \tag{12}$$

Proof. This is, again, a direct result of a basic property of conditional distributions:

$$\begin{aligned} D\left(p(x,y)\|q(x,y)\right) &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{q(x,y)}\right] \\ &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)p(x)}{q(y|x)q(x)}\right] \\ &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{q(y|x)}\right] + \mathbb{E}_{p(x,y)}\left[\log \frac{p(x)}{q(x)}\right] \\ &= D\left(p(y|x)\|q(y|x)\right) + D\left(p(x)\|q(x)\right) \end{aligned}$$

Hence proved. ■

1.6 Jensen's Inequality

This inequality would be helpful in later sections. One may remember this from Expectation-Maximization Algorithm. First remember the notion of functional convexity.

1.6.1 Convex Functions

★ **Definition 8.** (**Convex Function**) A function $f(x)$ is said to be convex over an interval (a,b) if for any points $x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$, we must have that:

$$\boxed{f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)} \tag{13}$$

Remark. A function is **strictly convex** if above holds only for $\lambda = 0$ or 1

⊛ **Theorem 6.** If the function f has a second derivative that is non-negative (positive) over an interval, then the function is convex (strictly convex) over that interval. That is,

$$\boxed{\begin{aligned} &\text{If } f''(x) \geq 0 \;\forall\; x \in (a,b) \implies f \text{ is convex over } (a,b). \\ &\text{If } f''(x) > 0 \;\forall\; x \in (a,b) \implies f \text{ is strictly convex over } (a,b). \end{aligned}} \tag{14}$$

⊛ **Theorem 7.** (**Jensen's Inequality**) If f is a convex function and X is a random variable, then:

- We have the following inequality

$$\boxed{\mathbb{E}_{p(X)}\left[f(x)\right] \geq f\left(\mathbb{E}_{p(X)}\left[X\right]\right)} \tag{15}$$

- We also have that:

$$\boxed{f \text{ is strictly convex} \implies X = \mathbb{E}_{p(X)}\left[X\right] \text{ In Probability.}} \tag{16}$$

Proof. TRIVIAL, by induction. ■

---

*The following theorem is of **fundamental importance** in Information Theory!*

⊛ **Theorem 8.** (**Information Inequality**) Let $p(x)$ and $q(x)$ for $x \in \mathcal{X}$ be two probability mass functions. Then we have the following two results:

$$\boxed{\begin{aligned} D\left(p\|q\right) &\geq 0 \\ D\left(p\|q\right) = 0 &\iff p(x) = q(x)\ \forall\ x \in \mathcal{X}. \end{aligned}} \tag{17}$$

Proof. Consider two mass functions $p(x)$ and $q(x)$ defined on same support $\mathcal{X}$. Hence,

$$\begin{aligned} -D\left(p\|q\right) &= -\mathbb{E}_{p(X)}\left[\log \frac{p(x)}{q(x)}\right] \\ &= -\sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x\in\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_{x\in\mathcal{X}} p(x) \frac{q(x)}{p(x)} \\ &\leq \log \sum_{x\in\mathcal{X}} p(x) \\ &\leq \log 1 \\ &\leq 0 \\ D\left(p\|q\right) &\geq 0 \end{aligned}$$

The other part follows trivially. ■

→ **Corollary 2.** (**Non-negativity of Mutual Information**) For any two random variables X and Y, we have

$$\boxed{D\left(p(y|x)\|q(y|x)\right) \geq 0}$$

,

$$\boxed{D\left(p(y|x)\|q(y|x)\right) = 0 \iff p(y|x) = q(y|x)}$$

,

$$\boxed{I(X;Y|Z) \geq 0}$$

,

$$\boxed{I(X;Y|Z) = 0 \iff X|Z \text{ and } Y|Z \text{ are independent.}}$$

,

$$\boxed{I(X;Y) \geq 0}$$

,

$$\boxed{I(X;Y) = 0 \iff X \text{ and } Y \text{ are independent.}}$$

Proof. Follows directly from the definition of $I(X;Y)$, $I(X;Y|Z)$ and the Information inequality. ■

⊛ **Theorem 9.** For a random variable X with sample space $\mathcal{X}$, we have:

$$\boxed{\begin{aligned} H(X) &\leq \log|\mathcal{X}| \\ H(X) = \log|\mathcal{X}| &\iff X \sim \text{Unif}(\mathcal{X}) \end{aligned}} \tag{18}$$

Proof. Define $u(x)$ to be the uniform distribution over space $\mathcal{X}$ and $p(x)$ to be the actual probability distribution over $\mathcal{X}$. Hence,

$$\begin{aligned} D(p\|u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log u(x) \\ &= -H(X) + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} \quad (\because u(x) = \frac{1}{|\mathcal{X}|}.) \\ &= -H(X) + \log|\mathcal{X}| \sum_{x\in\mathcal{X}} p(x) \\ &= -H(X) + \log|\mathcal{X}|\,. \end{aligned}$$

If $p(x)$ is uniform itself then we have that $D(p\|u) = 0$, hence the second result. Since $D(p\|u) \geq 0$, hence $H(X) \leq \log|\mathcal{X}|$. ■

Uncertainty reduces with more information.

⊛ **Theorem 10.** For two random variables X and Y, we have

$$\boxed{\begin{aligned} H(X|Y) &\leq H(X) \\ H(X|Y) = H(X) &\iff X \text{ and } Y \text{ are independent.} \end{aligned}} \tag{19}$$

Proof. Direct consequence of Mutual Information and the corollary that $I(X;Y) \geq 0$. ■

Remark. Note that this theorem doesn't mean that the uncertainty about X would be reduced by just the knowledge of one instance of Y. Specifically, this means that $H(X|Y=y)$ may be higher, lower or equal to $H(X)$, but on average, $\sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) = H(X|Y) \leq H(X)$.

→ **Corollary 3.** Let $X_1, \ldots, X_n$ be drawn according to JPMF $p(x_1, \ldots, x_n)$. Then,

$$\boxed{H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)} \tag{20}$$

1.7 Log-Sum Inequality

Theorem 11. (**Log-Sum Inequality**) For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, we have:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{21}$$

where equality is true if and only if $\frac{a_i}{b_i} = \text{constant } \forall\ i = 1, \ldots, n$.

Proof. Direct consequence of strict convex nature of $x \log x$ & Jensen's Inequality. ■

→ **Corollary 4.** The following are immediate/trivial results of Log-Sum Inequality:

1. Relative entropy $D(p\|q)$ is convex in the pair (p, q).
2. Entropy $H(p)$ is a concave function of p.

Theorem 12. Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

---

2 Asymptotic Equipartition Property

AEP is the information theoretic analogue of *Law of Large Numbers* in statistics. It divides the set of all sequences into *typical set* and *non-typical set* where typical set contains all those sequences whose sample entropy is close to true entropy. Moreover, if any property is true for the typical set, then it would be true for a large number of samples *with high probability*.

First, remember the notion of *convergence* in statistics.

⋆ **Definition 9.** (**Convergence of Random Variables**) Given a sequence of random variables $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$ ***converges*** to a random variable X :

- **In Probability** if for all $\epsilon > 0$, $\Pr\{|X_n - X| > \epsilon\} \longrightarrow 0$.
- **In Mean Square** if $\mathbb{E}\left[(X_n - X)^2\right] \longrightarrow 0$.
- **Almost Surely** if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$

2.1 The AEP Theorem

⊛ **Theorem 13.** (**AEP Theorem**) If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables sampled from $p(x)$, then we have that:

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \longrightarrow H(X) \quad \textit{In Probability} \tag{22}$$

Proof. Since X_i's are independent, hence we have that the probability of the sequence, i.e. $p(X_1, X_2, \ldots, X_n)$, are just the product of individual probabilities, that is,

$$p(X_1, X_2, \ldots, X_n) = p(X_1)p(X_2)\ldots p(X_n)$$

Hence, we get that,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_i \log p(X_i)$$

Now, by weak law of large numbers,

$$\frac{1}{n}\sum_i X_i \longrightarrow \mathbb{E}_{p(x)}[X] \quad \text{In Probability}$$

Now we know that any function of independent random variables is itself independent, hence $\log p(X_1), \ldots, \log p(X_n)$ are also independent since $p(X_i)$'s are itself function of X_i. Therefore, we must have

$$-\frac{1}{n}\sum_i \log p(X_i) \longrightarrow -\mathbb{E}_{p(x)}[\log p(X)] = H(X) \quad \text{In Probability}$$

Hence Proved. ■

⋆ **Definition 10.** (**Typical Set**) The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property that:

$$\forall (x_1, x_2 \ldots, x_n) \in A_\epsilon^{(n)}, \text{ we have, } 2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{23}$$

2.2 Basic Properties of Typical Set

Theorem 14. We have the following properties of $A_\epsilon^{(n)}$:

- $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)} \implies H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(x) + \epsilon$
- $\Pr\left\{(X_1, \ldots, X_n) \in A_\epsilon^{(n)}\right\} > 1 - \epsilon$ for n sufficiently large.
- $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$
- $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$

Proof. First part is trivial to see from the definition of $A_\epsilon^{(n)}$.
For the second part, first note that for any sequence $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$, we have the following:

$$\begin{aligned} H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) &\leq H(X) + \epsilon \\ -\epsilon \leq -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) &\leq \epsilon \\ \epsilon \geq \left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right|. & \end{aligned}$$

But we know that from AEP Theorem that $-\frac{1}{n} \log p(x_1, \ldots, x_n) \to H(X)$ in probability. What that means is:

$$\Pr\left\{\left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right| > \epsilon\right\} \to 0$$

This implies the following:

$$\Pr\left\{\left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right| \leq \epsilon\right\} \to 1$$

And this implies that, from the Cauchy's definition of convergence, for all $\delta > 0$, $\exists N \in \mathbb{N}$ such that

$$\left|\Pr\left\{\left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right| \leq \epsilon\right\} - 1\right| < \delta \; \forall \; n \geq N$$

Since probability is always positive, hence we can write it as:

$$\Pr\left\{\left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H(X) \right| \leq \epsilon\right\} > 1 - \delta \; \forall \; n \geq N.$$

Hence, for a particular value of $\delta = \epsilon > 0$, we would have N such that above condition is satisfied for all $n \geq N$, proving the second part.
The third part is easy to see:

$$\begin{aligned} 1 &= \sum_{(x_1,\ldots,x_n) \in \mathcal{X}^n} p(x_1, \ldots, x_n) \\ &\geq \sum_{(x_1,\ldots,x_n) \in A_\epsilon^{(n)}} p(x_1, \ldots, x_n) \quad (\because A_\epsilon^{(n)} \subseteq \mathcal{X}^n) \\ &\geq \sum_{(x_1,\ldots,x_n) \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)+\epsilon)} \end{aligned}$$

Therefore, we must have

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)},$$

proving the third part.
To prove the fourth part, note the second property implies that :

$$\begin{aligned} 1-\epsilon &< \Pr\left\{A_\epsilon^{(n)}\right\} \\ &\leq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(X)+\epsilon)} \end{aligned}$$

from which, it follows that,

$$(1-\epsilon)2^{n(H(X)+\epsilon)} \leq \left|A_\epsilon^{(n)}\right|.$$

■

Remark. It's important to note the following from the definition and the Theorem:

1. The notion of typicality is only concerned with the probability of a sequence and not the actual sequence itself.
2. The second property means that a sequence $(x_1, \ldots, x_n)$ drawn from $\mathcal{X}^n$ has a probability close to 1 ($1-\epsilon$ to be precise) to be present in $A_\epsilon^{(n)}$(!)
3. Moreover, the number of elements in typical set is nearly 2^{nH}.

---

3 Random Processes

⋆ **Definition 11.** (**Stationary Random Process**) A random process is said to be stationary if:

$$\boxed{\Pr\{X_1 = x_1, \ldots, X_n = x_n\} = \Pr\{X_{k+1} = x_1, \ldots, X_{k+n} = x_n\} \ \forall n, k \text{ and } (x_1, \ldots, x_n) \in \mathcal{X}.} \tag{24}$$

Remark. An example of a stationary process is a Markov Process.

⋆ **Definition 12.** (**Markov Process**) A discrete Random Process is said to be a Markov Process if:

$$\boxed{\begin{aligned}\Pr\{X_{n+1}|X_n = x_n, \ldots, X_1 = x_1\} = &\Pr\{X_{n+1} = x_{n+1}|X_n = x_n\} \\ &\forall\ x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}.\end{aligned}} \tag{25}$$

Remark. A Markov process is said to be **Time Invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n. In other words:

$$\Pr\{X_{n+1} = x|X_n = y\} = \Pr\{X_2 = x|X_1 = y\} \text{ for all } x, y \in \mathcal{X}. \tag{26}$$

3.1 Entropy Rate

⋆ **Definition 13.** (**Entropy of a Random Sequence**) The entropy of the random sequence $\{X_n\}$ is defined by

$$\boxed{H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \text{ when the limit exists.}} \tag{27}$$

Remark. This definition attempts to measure the rate at which the entropy of the sequence grows with n. This is the *per-symbol entropy of the n random variables.*

⋆ **Definition 14.** (**Alternate Entropy of a Random Sequence**) One can alternately define entropy rate as

$$\boxed{H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \text{ when the limit exists.}} \tag{28}$$

Remark. This definition, however, defines entropy rate as *the conditional entropy of the last random variable given all the previous.*

The following theorem states that **the above two notions of entropy rate are same for stationary processes.**

⊛ **Theorem 15.** For a stationary random process, the limits

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \text{ and } \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1) \text{ exists}$$

and

$$\boxed{H(\mathcal{X}) = H'(\mathcal{X}).} \tag{29}$$

⊛ **Theorem 16.** For a stationary Markov Chain, the entropy rate is given by:

$$\boxed{\begin{aligned}H(\mathcal{X}) &= H'(\mathcal{X}) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}) \\ &= H(X_2|X_1)\end{aligned}} \tag{30}$$

---

4 Codes

★ **Definition 15.** (**Source Code**) A source code C for a random variable X is the following map:

$$\boxed{\begin{aligned} &C : \mathcal{X} \longrightarrow \mathcal{D}^* \\ &\quad \text{where, } \mathcal{X} \text{ is the range of } X \text{ and,} \\ &\quad \mathcal{D}^* \text{ is the set of finite-length strings of symbols from a } D\text{-ary alphabet.} \end{aligned}} \tag{31}$$

Note that for $x \in \mathcal{X}$:

- $C(x) \in \mathcal{D}^*$ is the code associated to x by the above map.
- $l(x)$ denote the length of the code $C(x)$.

★ **Definition 16.** (**Average Length of Code**) The average length of a source code C defined over the random variable X with distribution $p(x)$ is given by:

$$\boxed{L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)} \tag{32}$$

★ **Definition 17.** (**Non-Singular Code**) A source code $C : \mathcal{X} \to \mathcal{D}^*$ is called Non-Singular if the map C is injective. That is,

$$\boxed{C(x) = C(y) \implies x = y \text{ for } x, y \in \mathcal{X}.} \tag{33}$$

Remark. Note that Non-Singularity makes sure that each event described by the range of the random variable X is uniquely identified by the corresponding code. That is, no two events have the same code.

★ **Definition 18.** (**Code Extension**) The code extension C^* for a code C is given by the map $C^* : \mathcal{X}^* \longrightarrow \mathcal{D}^*$ which takes each finite length string of elements of $\mathcal{X}$ to the finite length string of elements of the D-ary alphabet:

$$\boxed{C^*(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n).} \tag{34}$$

The multiplication $C(x_1)C(x_2)$ means concatenation of the two D-ary strings

Remark. Note that $C(x_1 x_2 \dots x_n)$ and $C^*(x_1 x_2 \dots x_n)$ means the same thing after the definition. Moreover, even though the definition is motivated by obvious practical reasons, it's interesting to see that the reminiscence of an algebraic homomorphism can be seen via the above definition! The instructor however doesn't wish to go there for obvious pedagogical reasons.

★ **Definition 19.** (**Uniquely Decodable Code**) A source code is called uniquely decodable if it's extension is Non-Singular.

Remark. Another way of saying this is that a uniquely decodable code has only one possible source producing it. An example that is not a uniquely decodable code is$A - 00, B - 100, C - 001$, which has ambiguity when encoding AB and CA. Therefore this is not a uniquely decodable code.

★ **Definition 20.** (**Prefix-Free Code**) A code is called a prefix-free code if no code-word is a prefix of another code-word.

Remark. Such codes are also called *Instantaneous Codes* as one can *instantaneously* infer the event in a sequence of events as soon as the code-word of that event is matched anywhere in the whole code.

4.1 Kraft's Inequality

Due to the obvious ease of decoding with Prefix-Free codes, our central aim hence is to construct minimum length, Prefix-Free Codes. One important result to this aim is given by the following theorem which ***limits the set of all codeword lengths possible for Prefix-Free codes.***

Theorem 17. (**Kraft's Inequality**) For any Prefix-Free code over an alphabet of size D, the countably infinite code-word lengths $l_1, l_2, \ldots$ satisfies

$$\boxed{\sum_i D^{-l_i} \leq 1.} \tag{35}$$

Note that the code-word lengths $l_1, l_2, \ldots$ here correspond to the length of the source code of each of the symbols/events for whom the encoding is being done.

Remark. Conversely, if $l_1, l_2, \ldots$ are countable natural numbers which satisfies the above equation, then there exists a prefix-free code with these code-word lengths.

4.2 Optimal Codes

Kraft's Inequality showed us the necessary and sufficient conditions to guarantee the existence of Prefix-Free codes. We now deal with the problem of actually finding the Prefix-Free codes with minimum expected length.

Consider the following chain of statements:

1. In order to find the minimum length prefix-free codes, by Kraft's Inequality, it is sufficient to find natural numbers $l_1, \ldots, l_m$ satisfying Kraft's Inequality and that the average expected length is the least amongst the other possible prefix-free codes.

2. This hence becomes the following minimization problem:

$$\begin{aligned} &\text{Minimize : } L = \sum_i p_i l_i \text{ over all integers } l_1, l_2, \ldots l_m \\ &\text{satisfying,} \\ &\sum_i D^{-l_i} \leq 1 \end{aligned} \tag{36}$$

3. Using the general method of *Lagrange Multipliers*, we transform the above Minimization problem into the following equation (note that we remove the integer constraint on l_i's and consider the inequality as an equality):

$$J = \sum_i p_i l_i + \lambda \sum_i D^{-l_i}$$

   Then the stationary points of J can be found by setting derivative to zero:

$$\begin{aligned} \frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \ln D = 0 \\ \implies D^{-l_i} = \frac{p_i}{\lambda \ln D} \end{aligned}$$

4. Substituting $D^{-l_i} = \frac{p_i}{\lambda \ln D}$ in the Minimization constraint $\sum_i D^{-l_i} = 1$, we see

$$\begin{aligned} \frac{1}{\lambda \ln D} \sum_i p_i = 1 \\ \frac{1}{\ln D} = \lambda \end{aligned}$$

   Substituting this λ in $D^{-l_i} = \frac{p_i}{\lambda \ln D}$, we get

$$\begin{aligned} p_i &= D^{-l_i} \\ \implies \quad l_i &= -\log_D p_i \end{aligned} \tag{37}$$

   which is the required condition for minimum average length Prefix-Free codes!