- Lecture 2:

Markov Chain

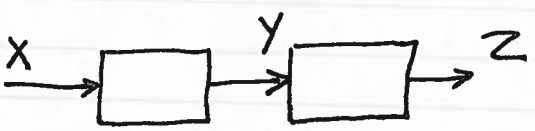
Random variables X, Y, Z form a Markov Chain if the conditional distribution of Z depends only on Y and is conditionally independent of X. That is if one knows the value of Y then X cannot say anything more about Z.
We show this as $X \to Y \to Z$ and formally, we have $X \to Y \to Z$ if

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

if $X \to Y \to Z$ then X and Z are conditionally independent, i.e.,

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

Data Processing Inequality



If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$

Proof:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$$

$I(X;Z|Y) = 0$ since X and Z are conditionally independent.

So,

$$I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Z)$$

In particular,

if $Z = g(Y)$, then:

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

since

$$X \to Y \to g(Y)$$

Corollary: If $X \to Y \to Z$ then:

$$I(X;Y|Z) \leq I(X;Y)$$

Proof:

$$I(X;Y,Z) = I(X;Y) + \cancel{I(X;Z|Y)}^{0}$$

$$= I(X;Z) + I(X;Y|Z)$$

$$I(X;Y|Z) = I(X;Y) - I(X;Z) \leq I(X;Y)$$

Sufficient Statistics

Assume that we have a family probability mass function $\{f_\theta(x)\}$ indexed by θ. Let X be a sample from a distribution in this family. Take $T(X)$ to be any statistic (function of the sample)

like sample mean, sample variance, etc.

Then $\theta \to X \to T(X)$ and by data processing theorem:

$$I(\theta; T(X)) \leq I(\theta; X)$$

A statistic $T(X)$ is called <u>sufficient</u> if the above holds with equality.

Definition: A function $T(X)$ is a <u>sufficient statistic</u> if X is independent of θ given $T(X)$, i.e., $\theta \to T(X) \to X$. This is equivalent to:

$$I(\theta; X) = I(\theta; T(X))$$

for all distributions on θ.

<u>Example</u>:

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta)^2}{2}} = N(\theta, 0)$$

Let $X_1, \ldots X_n$ be drawn independently according to this distribution. The the suffilient statistics for θ is $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. It can be shown that the conditional distribution of $X_1, \ldots X_n$ given on $\overline{X_n}$ and n does not depend on θ.

Fano's inequality

Fano's inequality relates the probability of error and Channel equivocation.

That is $P_e = P_r(\hat{X} \neq X)$ and $H(X|Y)$

Suppose we wish to estimate X based on the observation of Y. X is distributed as $p(x)$ and Y is related to X through $p(y|x)$, then:

Theorem:

$$H(P_e) + P_e \log(|X|-1) \geq H(X|Y)$$

a little bit weaker version is

$$1 + P_e \log|X| \geq H(X|Y)$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log|X|}$$

Proof:

Define r.v.,

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$$

Using the chain rule, we have:

$$H(E,X|Y) = H(X|Y) + \cancel{H(E|X,Y)}^{\,0}$$

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

$$H(E|Y) \leq H(E) = H(P_e)$$

$$H(X|E,Y) = Pr(E=0)\,H(X|Y,E=0) + Pr(E=1)\,H(X|Y,E=1)$$

$$\leq (1-P_e)\,0 + P_e \log(|X|-1)$$

So,

$$H(X|Y) \leq H(P_e) + P_e \log(|X|-1)$$

The Asymptotic Equipartion Property (AEP)

AEP is the application of the law of large numbers in information theory. The same way that $\frac{1}{n}\sum_{n=1}^{n} X_i \to E[X]$ we show that

$$-\frac{1}{n}\log p(X_1, \cdots X_n) \to H(X)$$

This means that $p(X_1, \cdots X_n)$ will be close to $2^{-nH(X)}$.

Theorem (AEP): If $X_1, \cdots X_n$ are i.i.d. $\sim p(x)$ then

$$-\frac{1}{n}\log p(X_1, \cdots, X_n) \to H(X) \text{ in probability}$$

that is for any $\varepsilon > 0$ there is some n_0 such that for any $n > n_0$

$$\Pr\left[\left|-\frac{1}{n}\log p(X_1, \cdots X_n) - H(X)\right| > \varepsilon\right] < \delta$$

Proof:

$$-\frac{1}{n}\log p(X_1, \cdots X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i)$$

$$\to -E[\log p(X_i)] \quad \leftarrow \text{(using weak law of large numbers)}$$

$$= H(X)$$

Definition The typical set $A_\varepsilon^{(n)}$ (ε-typical set) with respect to $p(x)$ is defined as the set of all sequences $(x_1, x_2, \ldots x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, \ldots x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

Theorem: The typical set $A_\varepsilon^{(n)}$ has the following properties:

1) for each $(x_1, \ldots x_n) \in A_\varepsilon^{(n)}$

$$H(X) - \varepsilon \leq -\frac{1}{n} \log p(x_1, \ldots x_n) \leq H(X) + \varepsilon$$

2) $P_r\{A_\varepsilon^{(n)}\} > 1 - \varepsilon$ for sufficiently large n.

3) $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$ where $|A|$ denotes the cardinality of A.

4) $|A_\varepsilon^{(n)}| \geq (1-\varepsilon) 2^{n(H(X)-\varepsilon)}$

Proof:

1) from the definition

$$-n(H(X)+\varepsilon) \leq \log p(x_1 \ldots x_n) \leq -n(H(X)-\varepsilon)$$

$$H(X) - \varepsilon \leq -\frac{1}{n} \log p(x_1 \ldots x_n) \leq H(X) + \varepsilon$$

2) For sufficiently large n

$$Pr\{A_\varepsilon^{(n)}\} = Pr\{(x_1,\dots,x_n) \in A_\varepsilon^{(n)}\}$$

$$= P\left\{\left|-\frac{1}{n}\log p(x_1,\dots,x_n) - H(X)\right| < \varepsilon\right\}$$

$$= 1 - P\left\{\left|-\frac{1}{n}\log p(x_1\cdots x_n) - H(X)\right| > \varepsilon\right\} \geq 1-\delta$$

by choosing $\delta = \varepsilon$ part 2 is proven.

3)

$$Pr\{A_\varepsilon^{(n)}\} = \sum_{A_\varepsilon^{(n)}} p(\underline{x}) \leq \sum_{X^n} p(\underline{x}) = 1$$

So

$$1 \geq \sum_{A_\varepsilon^{(n)}} p(\underline{x}) \geq \sum_{A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)} |A_\varepsilon^{(n)}|$$

and, therefore,

$$|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$$

4) $Pr(A_\varepsilon^{(n)}) \geq 1-\varepsilon$ from property 2

So

$$1-\varepsilon \leq Pr(A_\varepsilon^{(n)}) = \sum_{A_\varepsilon^{(n)}} p(\underline{x}) \leq \sum_{A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)} |A_\varepsilon^{(n)}|$$

Therefore:

$$|A_\varepsilon^{(n)}| \geq (1-\varepsilon)\, 2^{-n(H(X)+\varepsilon)}$$

in summary:

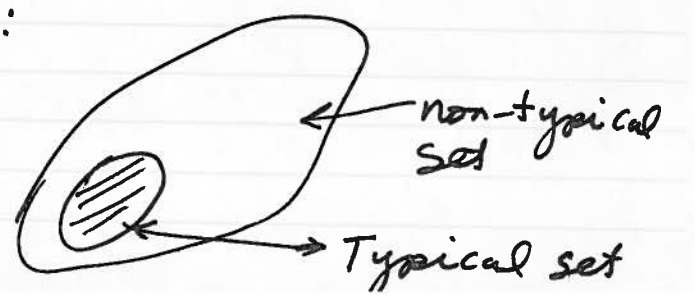
$$(1-\varepsilon)\, 2^{-n(H(X)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{-n(H(X)+\varepsilon)}$$

Interpretation:

The AEP indicates that the number of realistic possibilities of a sequence of i.i.d. events is, usually, much less than the $|\mathcal{X}|^n$ possibilities. It also indicates that all those possibilities that usually occur have roughly the same probability. In essence, it reveals the "dullness" of the world in the sense that first of all, out of a huge number of possibilities only just a few happen in practice and secondly, those occuring roughly cause the same level of "surprise".

Application of AEP in data Compression:

Assume that $X_1, \dots X_n$ are $\sim p(x)$, i.i.d. and we wish to represent them. We divide the total $|\mathcal{X}|^n$ sequences into the set of typical sequences $A_\varepsilon^{(n)}$ and having less than $2^{n(H(X)+\varepsilon)}$ members and the set of non-typical sequences.



To represent each typical sequence, we need at most $\log 2^{n(H(X)+\epsilon)} + 1 = n(H(X)+\epsilon)+1$ bits. We devise a coding scheme as follows:

for each sequence, we specify the sequence to be a typical or non-typical by adding preumble bit, say: 0 for typical and
1 for non-typical

then, we use:

1) $n(H(X)+\epsilon)+1$ bits for each typical sequence. We can use any arbitrary indexing for counting the $2^{n(H(X)+\epsilon)}$ typical sequences.

2) $n \log|X|+1$ ~~per~~ per non-typical sequence.

So:

$$E[\ell(X^n)] = \sum_{x^n} p(x^n)\ell(x^n) = \sum_{x^n \in A_\epsilon^{(n)}} p(x_n)\ell(x_n^n) + \sum_{x^n \notin A_\epsilon^{(n)}} p(x^n)\ell(x^n)$$

$$\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)[n(H+\epsilon)+2] + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n)(n\log|X|+2)$$

$$\leq n(H+\epsilon) + \epsilon n \log|X| + 2$$

$$= n(H + \underbrace{\epsilon + \epsilon \log|X| + \frac{2}{n}}_{\epsilon'})$$

The extra term (summed into ε') can be made arbitrarily small by proper choice of ε and n.

Theorem: Let X^n be i.i.d. $\sim p(x)$. Let $\varepsilon > 0$. Then, there is a code that maps sequences x^n of length n into binary strings such that the mapping is one-to-one, therefore, lossless and

$$E\left[\frac{1}{n} l(x^n)\right] \leq H(X) + \varepsilon$$

Entropy Rate of a Stochastic Process

In order to extend the notion of entropy, which we have only defined for i.i.d. sources to sources with memory, we define the entropy rate:

Definition: The entropy rate of a random process X is defineds:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n)$$

Another definition that we show later to be equivalent to the above is given as

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \cdots X_1)$$

before proving the equivalence of $H(\mathcal{X})$ and $H'(\mathcal{X})$ let's find the entropy rate of the memoryless processes.

For an i.i.d. process:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, \cdots, X_n)}{n} = \lim_{n\to\infty} \frac{nH(X)}{n} = H(X_1)$$

Similarly, for a source that is memoryless but not with identically distributed samples, we have:

$$H(X_1, \cdots X_n) = \sum_{i=1}^{n} H(X_i)$$

In this case, it is possible that $\lim_{n\to\infty} \frac{H(X_1, \cdots X_n)}{n}$ does not exist.

As an example, take a binary sequence with

$$p_i = P(X_i = 1) = \begin{cases} 0.5 & 2k \le \log\log i \le 2k+1 \\ 0 & 2k+1 \le \log\log i \le 2k+2 \end{cases}$$

for $k = 0, 1, 2, \cdots$

In this example $H(X_i)$ will be 1 for certain period of time followed by a longer period of time when it $H(X_i) = 0$. So, the running average oscillates between 0 and 1.

To show that $H(\mathcal{X}) = H'(\mathcal{X})$, we first prove that the limit $H'(\mathcal{X}) = \lim_{n\to\infty} (H(X_n|X_{n-1},\dots X_1))$ exists.

Theorem: For a stationary stochastic Process, $H(X_n|X_{n-1},\dots,X_1)$ is decreasing in n and has limit $H'(\mathcal{X})$.

Proof:

$$H(X_{n+1}|X_1,\dots,X_n) \leq H(X_{n+1}|X_n,\dots,X_2)$$

$$= H(X_n|X_{n-1},\dots,X_1)$$

So $H(X_n|X_{n-1},\dots,X_1)$ is a decreasing sequence of non-negative number and, therefore, must have a limit $H'(\mathcal{X})$.

Another theorem that we need in order to prove the equivalence of $H(\mathcal{X})$ and $H'(\mathcal{X})$ is:

Theorem: If $a_n \to a$ then $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i \to a$.

Proof: since $a_n \to a$, there is some $N(\varepsilon)$ s.t., $|a_n - a| < \varepsilon$, $n \geq N$

$$|b_n - a| = \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \leq \frac{1}{n}\sum_{i=1}^{n}|a_i - a|$$

$$= \frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|a_i - a| + \frac{1}{n}\sum_{i=N(\varepsilon)}^{n}|a_i - a|$$

$$\leq \frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|a_i - a| + \frac{n - N(\varepsilon)}{n}\varepsilon \qquad \forall\, n \geq N(\varepsilon)$$

as $n \to \infty$, the first term $\to 0$ so $b_n \to a$.

Now, we are ready to prove the equivalence of $H(\mathcal{X})$ and $H'(\mathcal{X})$.

Theorem: For a stationary stochastic process

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, \cdots X_n)}{n} = \lim_{n\to\infty} H(X_n | X_{n-1}, \cdots X_1) = H'(\mathcal{X})$$

Proof:

by the chain rule

$$\frac{1}{n} H(X_1, \cdots X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1)$$

So:

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots X_n) = \lim_{n\to\infty} H(X_n | X_{n-1}, \cdots X_1) = H'(\mathcal{X}).$$

Here, we have used the fact that the limit of ~~each~~ the terms in the summation is $H'(\mathcal{X})$.

The significance of the entropy rate is due to the AEP property:

$$-\frac{1}{n} \log p(X_1, \cdots, X_n) \to H(\mathcal{X})$$

with probability 1.

Definition: #(Markov Chain): A random process $X_1, X_2, \cdots X_n, \cdots$ is said to form a Markov Chain if: