August 30th, 2017.

Lec7: Asymptotic Equipartition property (AEP)

Recall:
- Data processing inequality
- Sufficient statistics
- Fano's Inequality

Today
Begin with background.

Today
- AEP
- Consequences

Convergence of RVs

Let $\{X_n\}_{n\geq 1}$ be a sequence of RVs. Then (a) X_n converges in the mean square sense (MSS) if $\lim_{n\to\infty} E\left[|X_n - X|^2\right] = 0$

(b) $\{X_n\}$ converges w.p. 1 or almost everywhere (surely) (a.e) if

$$\Pr\{w : \lim_{n\to\infty} X_n(w) = X(w)\} = 1$$

(c) $\{X_n\}$ converges to X in probability if given any $\epsilon > 0$.

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0$$

i.e., Given any $\epsilon, \delta > 0$, $\exists\, n_{\epsilon,\delta}$ such that $\forall\, n \geq n_{\epsilon,\delta}$

$$\Pr(|X_n - X| > \epsilon) < \delta$$

Can be shown that

MSS $\Rightarrow$ convergence in probability

conv a.e $\Rightarrow$

Markov inequality

$$\Pr\{X \geq a\} \leq \frac{E[X]}{a}$$

where X is a non-negative random variable, $a > 0$.

$$E[X] = \sum_{x\geq 0} x\, p(x) \geq \sum_{x\geq a} x\, p(x) \geq a\, P(X \geq a)$$

Chebychev inequality

$$\Pr\{|X - \bar{X}| > a\} \leq \frac{E\{|X - \bar{X}|^2\}}{a^2} = \frac{Var(X)}{a^2} = \frac{\sigma^2}{a^2}.$$

Weak law of large numbers:

Let $\{X_n\}$ be a sequence of iid random variables having common probability mass function $p(x)$. $E[X] = \sum_{x\in\mathcal{X}} x\, p(x)$

Set $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow$ {Successive sample means}

(Assume variance $\sigma^2 < \infty$)

Claim: $Y_n \to E[X]$ in probability.

Pf: $\Pr\{|Y_n - E[X]| > \epsilon\} \leq \frac{E\left[|Y_n - E[X]|^2\right]}{\epsilon^2}$

$$= \frac{E\left[\frac{1}{n^2}\left(\sum_{i=1}^{n}(X_i - E[X])\right)^2\right]}{\epsilon^2} = \frac{1}{n^2\epsilon^2} E\left[\left(\sum_{i=1}^{n}(X_i - E[X])\right)^2\right]$$

$$E\left[\left(\sum_{i=1}^{n} x_i - E[x]\right)^2\right] = E\left[\sum_i (x_i - E[x])^2\right] + E\Big[\underbrace{\sum_{i \neq j} (x_i - E[x])(x_j - E[x])}_{\substack{\downarrow \\ 0 \text{ as } x_i, x_j \text{ independant}}}\Big]$$

cross terms vanish

$$\therefore P\left(|Y_n - E[x]| > \epsilon\right) \leq \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

where $\sigma^2 = E[(x_i - E[x])^2]$

$$\therefore \lim_{n \to \infty} P\left(|Y_n - E[x]| > \epsilon\right) = 0$$

for given any ϵ, δ can find large enough n such that $\frac{\sigma^2}{n\epsilon^2} < \delta$.

Corollary: If $\{X_n\}$ is iid with probability mass function $p(x)$ then alphabet $|\mathcal{X}| < \infty$

$$\frac{1}{n} \log \frac{1}{P(x_1, \ldots x_n)} \longrightarrow H(X) \text{ in probability.}$$

Pf: Set $Y_i = \frac{1}{\cancel{\log}} \log\left(\frac{1}{P(X_i)}\right)$ then Y_i are also iid.

$$\frac{1}{n} \log \frac{1}{P(x_1, \ldots x_n)} = \frac{\sum_{i=1}^{n} Y_i}{n} \xrightarrow[\text{In probability}]{\text{convergence}} E[Y] = E\left[\log \frac{1}{P(X)}\right] = \underline{\underline{H(X)}}$$

AEP (Asymptotic Equipartition property)

(Motivation)

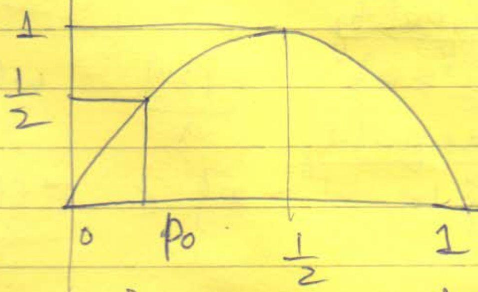
$X \in \{0, 1\}$.

$$X = \begin{cases} 1 & \text{prob } p \\ 0 & \text{prob } 1-p \end{cases}$$

(Bernoulli RV)

What is $H(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{(1-p)}$

$p^{NP} (1-p)^{N(1-p)}$



i.e., $H(p_0) = \frac{1}{2}$ for some $p_0 < \frac{1}{2}$

$$\underline{X} = \{X_n\}_{n=1}^{N} \longrightarrow \binom{N}{2} \leftarrow \text{uncertainity in } \underline{X}$$

$\downarrow$

X_n iid

$X_n \sim (1-P_0, P_0)$ (0, 1) $\qquad P_0 : H(P_0) = \frac{1}{2}$.

Most likely sequence of length N.

$\underline{0} \Rightarrow \text{prob} = (1-p)^N$.

$$\frac{1}{N} \log \frac{1}{(1-p)^N} = \log \frac{1}{1-p}.$$

doesn't look like entropy $H(p)$

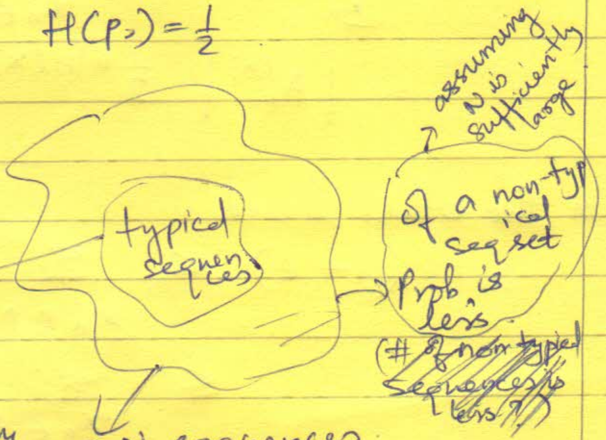
Consider a sequence with 'k' 1's where $\frac{k}{N} \approx p_0$ then probability of this sequence $= (p_0)^{k} (1-p_0)^{N-k}$

$$= (p_0)^{Np_0} (1-p_0)^{N(1-p_0)}$$

$$\frac{1}{N} \log \frac{1}{(P_0)^{NP_0}(1-P_0)^{N(1-P_0)}} = P_0 \log \frac{1}{P_0} + (1-P_0) \log \frac{1}{1-P_0} = \frac{1}{2}$$

looks like the uncertainity for this sequence is $H(P_0) = \frac{1}{2}$

~~We call~~ these sequences are "typical" sequences.



$$2^{N/2} = 2^{N H(P_0)} \simeq \binom{N}{K} = \binom{N}{NP_0}$$

Consider that there only 2^N sequences typical sequences to be represented, one can do that using $\frac{N}{2}$ bits in this case.

<u>Setting</u>: $\{X_n\}$ iid with common pmf $p(x)$. Define for $\epsilon > 0$, $n \geq 1$.

$$A_\epsilon^{(n)} = \left\{ \underline{x}_n \;\middle|\; H(x) - \epsilon \leq \frac{1}{n} \log \frac{1}{P(\underline{x}_n)} \leq H(x) + \epsilon \right\} \rightarrow (1)$$

Call this a typical set

<u>Note</u>: $\underline{x} \in A_\epsilon^{(n)}$

$$\Rightarrow n(H(x) - \epsilon) \leq -\log p(\underline{x}) \leq n(H(x) + \epsilon)$$

$$\Rightarrow 2^{-n(H(x)+\epsilon)} \leq p(\underline{x}) \leq 2^{-n(H(x)-\epsilon)} \rightarrow (2)$$

<u>Claim</u>: $\exists\, n_\epsilon$ such that $\forall\, n > n_\epsilon$.

$$\Pr\{\underline{X}_n \in A_\epsilon^{(n)}\} \geq 1 - \epsilon \qquad \underline{X}_n \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

<u>Proof</u>:

We know that

$$\frac{1}{n} \log \frac{1}{P(\underline{X}_n)} \longrightarrow H(X) \text{ in prob.}$$

$\therefore$ for given $\epsilon > 0$, $\delta > 0$, $\exists\, n_{\delta,\epsilon}$ such that $\forall\, n > n_{\delta,\epsilon}$

$$\Pr\left\{ \left| \frac{1}{n} \log \frac{1}{P(\underline{X}_n)} - H(x) \right| > \epsilon \right\} \leq \delta$$

$$\Longleftrightarrow \Pr\left\{ \underline{X}_n \in A_\epsilon^{(n)} \right\} \geq 1 - \delta$$

Choosing $\delta < \epsilon$ we obtain

that $\Pr\{\underline{X}_n \in A_\epsilon^{(n)}\} \geq 1 - \epsilon \quad \forall\, n > n_{\delta,\epsilon}$

Size of $A_\epsilon^{(n)}$

$$1 = \sum_{x_n} p(x_n) \geq \sum_{x_n \in A_\epsilon^{(n)}} p(x_n) \geq \sum_{x_n \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \rightarrow \text{(using ②)}$$

$$= |A_\epsilon^{(n)}| \, 2^{-n(H(X)+\epsilon)} \Longrightarrow |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$$

On the other hand

$$(1-\epsilon) < \sum_{x_n \in A_\epsilon^{(n)}} p(x_n)$$ given we picked large enough $n \geq n_\epsilon$

$$\leq \sum_{x_n \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)}$$ (using ②)

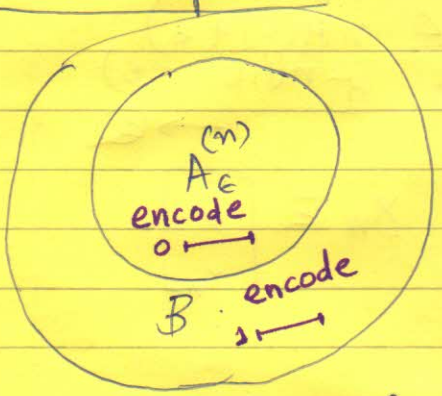
$$= |A_\epsilon^{(n)}| \, 2^{-n(H(X)-\epsilon)}$$

$$\Rightarrow \boxed{A_\epsilon^{(n)} > (1-\epsilon)\, 2^{n(H(X)-\epsilon)}} \rightarrow ③$$

Hence we obtain

$$\boxed{(1-\epsilon)\, 2^{n(H(X)-\epsilon)} < A_\epsilon \leq 2^{n(H(X)+\epsilon)}}$$

Data Compression



$$B = \left(A_\epsilon^{(n)}\right)^c$$

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)} \leq 2^t \quad \text{for } t \leq n(H(X)+\epsilon)+1$$

$$|B| \leq |\mathcal{X}|^n \leq 2^s \quad \text{for } s \leq n\log|\mathcal{X}| + 1$$

$$\therefore \text{Average length} \leq t \underbrace{P\{X_n \in A_\epsilon^{(n)}\}}_{\leq 1} + s\, P\{X_n \in B\} + \underbrace{1}_{\text{for } 0,1 \text{ precoding}}$$

$$l \leq (nH(x) + n\epsilon + 1) + (n\log|\mathcal{X}| + 1)\epsilon + 1$$

$$= nH(x) + n\epsilon + (\epsilon + 2) + n\log|\mathcal{X}|\,\epsilon$$

$$\frac{l}{n} \leq H(x) + \epsilon\log|\mathcal{X}| + \epsilon + \left(\frac{\epsilon+2}{n}\right)$$

Which can be made smaller than any $\delta > 0$ by the choice of n, ϵ. then

$$\frac{l}{n} \leq H(x) + \delta.$$

Thm: Let $\{X_n\}$ be iid $\sim p(x)$. Let $\delta > 0$, then $\exists$ a code mapping strings $\underline{x}_n$ to l-bit binary strings st the mapping is one to one and $\mathbb{E}\left\{\frac{1}{n} l(\underline{x}_n)\right\} \leq H(X) + \delta$ //

Same setting

Let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ st $P_r(B_\delta^{(n)}) \geq 1-\delta$, then $B_\delta^{(n)}$ must have sufficient intersections with $A_\epsilon^{(n)}$.

$$P_r(\underline{X}_n \in A_\epsilon^{(n)}) > 1-\epsilon$$

$$P_r(\underline{X}_n \in B_\delta^{(n)}) \geq 1-\delta$$

$$P_r(\underline{X}_n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}) + P(\underline{X}_n \in A_\epsilon^{(n)} \cup B_\delta^{(n)}) \geq 2-\epsilon-\delta$$

$$\Rightarrow P_r(\underline{X}_n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 2-\epsilon-\delta - P(\underline{X}_n \in A_\epsilon^{(n)} \cup B_\delta^{(n)})$$
$$> 1-\epsilon-\delta.$$

$$1-\epsilon-\delta < \sum_{\underline{x}_n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} P(\underline{x}_n)$$
$$\leq |A_\epsilon^{(n)} \cap B_\delta^{(n)}| \, 2^{-n(H(X)-\epsilon)}$$
$$\Rightarrow |B_\delta^{(n)}| \geq |A_\epsilon^{(n)} \cap B_\delta^{(n)}| \geq (1-\epsilon-\delta)\, 2^{n(H(X)-\epsilon)}$$

In other words any other subset with high probability must have significant intersection with the typical set.