Lecture 10: Information Theory & Digital Communication,

Capacity of Binary Erasure Channel (BEC)

Last time: modes of convergence

Ex: convergence a.s., in prob, in dist

WLLN: majority of people are getting closer to true mea

Ex: WLLN is an example of convergence in probability

SLLN: tells us the individual sample path $= E[X]$ as $n \to \infty$ w.p. 1.

Ex: SLLN is an example of convergence almost surely

Central Limit Thm

Ex: CLT is an example of convergence in distribution

$$(X_n)_{n \geq 1} \underset{\text{iid}}{\sim} X$$

how empirical mean varies from true mean

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{(X_i - \mu)}{\sigma}} \to Z \sim \mathcal{N}(0,1) \text{ in dist.}$$

"Fluctuations of avg around true avg is std normal in distribution"

MGFs of sequence of Y_n's

"Pf": "Fact": If $\underbrace{M_{Y_n}}(t) \to M_Y(t)$ $\forall t \in \mathbb{R} \Rightarrow Y_n \to Y$ in distribution

Now, WLOG, assume $\mu = 0$, $\sigma = 1$

We can write:

$$M_{X_1}(t) = E[e^{tX}] = \sum_{n \geq 0} t^n \frac{E[X^n]}{n!} = 1 + \frac{t^2}{2} + \underbrace{o(t^2)}_{\text{smaller order terms}}$$

goes $\to 0$ faster than $t \to 0$

bc iid   asymptotic expansion

$$M_{\frac{X_1 + \cdots + X_n}{\sqrt{n}}}(t) = \left(M_{X_1}\left(\frac{t}{\sqrt{n}}\right) \right)^n = \left[M_{X_1}\left(\frac{t}{\sqrt{n}}\right) \right]^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{t}{n}\right)\right)^n$$

$$\to e^{t^2/2} = \underbrace{M_Z(t)}_{\text{std normal}} \qquad \lim_{n \to \infty} \left(1 + \frac{c}{n}\right)^n = e^c$$

By "Fact" CLT follows

There are many other modes of convergence

ex: $X_n \to X$ in L^p, $p \geq 1$ if $E|X_n - X|^p \to 0$

by Markov's: convergence in L^p $\Rightarrow$ convergence in prob $\Rightarrow$ conv in distribution

$\not\Downarrow$ $\not\Uparrow$ $\nearrow$

convergence a.s.

Information Theory:

↳ how probability can be applied to complex problems

Questions Shannon addressed:

1. How reliably can I transmit information from point A to point B? And at what rate can I do so?
   [channel coding problem]

   ex: cocktail party problem
   ↳ hard to hear someone at a party:
   → talk louder (↑ power), repeat self (↓ rate), add context

2. How many bits on average do I need to describe the outcome of an experiment? [source coding problem]
   ↳ depending on how random an experiment is, might need more/less bits

   ex: data compression
   ↳ can do predictive coding

Source Coding (ie, Compression)

For discrete rv $\mathcal{X}$ with PMF P_x ($\mathcal{X} \sim P_x$) we define

Shannon Entropy:

$$H(x) := \mathbb{E}\left[\log\left(\frac{1}{P_x(\mathcal{X})}\right)\right]$$

$$= \sum_x P_x(x) \log\left(\frac{1}{P_x(x)}\right)$$

$\log \equiv \log_2$

↑ can use any base but base 2 tells us that units of entropy are bits

intuitively, entropy $H(x)$ is how random $\mathcal{X}$ is or uncertainty about $\mathcal{X}$ on average

precisely, from source coding theorem:

$H(\mathcal{X}) :=$ # of bits needed to describe $\mathcal{X}$, on avg. "number of Yes/No questions needed to determine $\mathcal{X}$"

Source Coding Theorem: For any $\varepsilon > 0$ and $(X_i)_{i \geq 1} \underset{\text{iid}}{\sim} X$

the vector $(X_1, \dots, X_n)$ can be represented using

$\leq n(H(X) + \varepsilon)$ bits on average.

Moreover, representation $< nH(X)$ incurs loss.