LECTURE 4

Convergence and Asymptotic Equipartition Property

Last time:

- Fano's Inequality
- Stochastic Processes
- Entropy Rate
- Hiden Markov Process

Lecture outline

- Types of convergence
- Weak Law of Large Numbers
- Strong Law of Large Numbers
- Asymptotic Equipartition Property

Reading: Chapter 3.

Convergence of Random Variables

A sequence of maps $\Omega \to \mathcal{X}$ converge, w.o.l.g., to 0.

Pointwise convergence: for any $\omega \in \Omega$, $X_n(\omega) \to 0$.

Goal The Law of Large Numbers: the average of a sequence of i.i.d. r.v.s converges to the mean.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_n \to E[X]$$

Need weaker notions.

Types of convergence

- Almost sure convergence (also called convergence with probability 1)

$$P\left(\left\{\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\}\right) = 1$$

write $Y_n \to Y \quad a.s.$.

- Mean-square convergence:

$$\lim_{n\to\infty} E[|Y_n - Y|^2] = 0$$

- Convergence in probability: $\forall \epsilon > 0$

$$\lim_{n\to\infty} P\left(\{\omega : |Y_n(\omega) - Y(\omega)| > \epsilon\}\right) = 0$$

- Convergence in distribution: the cumulative distribution function (CDF) $F_n(y) = Pr(Y_n \leq y)$ satisfy

$$\lim_{n\to\infty} F_n(y) \to F_Y(y)$$

  at all y for which F is continuous.

Relations among types of convergence

Venn diagram of relation:

Weak Law of Large Numbers

$X_1, X_2, \ldots$ i.i.d.
finite mean μ and variance σ^2

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

- $\mathbf{E}[S_n] =$

- $\text{Var}(S_n) =$

- As n increases, S_n is distributed around μ with a smaller variance.
- Smaller variance means S_n cannot be too far away from its mean —— need to make rigorous.

Chebyshev's Inequality

Theorem Consider random variable Z taking on only nonnegative values, $\forall \delta > 0$,

$$P(Z \geq \delta) \leq \frac{1}{\delta} E[Z]$$

Proof

$$\begin{aligned} E[Z] &= P(Z \geq \delta)E[Z|Z \geq \delta] + P(Z < \delta)E[Z|Z < \delta] \\ &\geq P(Z \geq \delta)E[Z|Z \geq \delta] \\ &\geq P(Z \geq \delta)\delta \end{aligned}$$

Let S be a zero mean r.v. with variance σ_S^2, let $Z = S^2 \geq 0$. $E[Z] = \sigma_S^2$.

Apply Chebyshev's inequality,

$$P(|S| \geq k\sigma_S) = P(Z \geq k^2\sigma_S^2) \leq \frac{1}{k^2}$$

Finishing the Proof of the Weak LLN

Recall $S_n = \frac{1}{n}(X_1 + \ldots + X_n)$, with $E[S_n] = \mu$, and $\text{Var}[S_n] = \frac{\sigma^2}{n}$, we have

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_n - \mu\right| \geq \frac{k\sigma}{\sqrt{n}}\right] \leq \frac{1}{k^2}$$

For any ϵ, take large n and k, let $\epsilon = \frac{k\sigma}{\sqrt{n}}$.

AEP

If $X_1, \ldots, X_n$ are IID with distribution P_X, then

$-\frac{1}{n}\log(P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)) \to H(X)$ in probability

Proof: create r.v. $Y = \log(P_X(X))$: i.e. Y takes the value $y_i = \log(P_X(x_i))$ with probability $P_X(x_i)$ (note that the value of Y is related to its probability distribution)

we now apply the WLLN to Y

AEP

For any $\omega \in \Omega$,

$$
\begin{aligned}
& -\frac{1}{n}\log(P(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))) \\
= & -\frac{1}{n}\sum_{i=1}^{n} P_X(X_i(\omega)) \\
= & -\frac{1}{n}\sum_{i=1}^{n} Y_i(\omega)
\end{aligned}
$$

using the WLLN on Y

$-\frac{1}{n}\sum_{i=1}^{n} Y_i \to E_Y[Y]$ in probability, i.e., $\forall \epsilon$,

$$\lim_{n\to\infty} P\left[\left|-\frac{1}{n}\sum_{i=1}^{n}\log P_X(X_i) - E[Y]\right| \le \epsilon\right] = 1$$

$E[Y] = -E[\log(P_X(X))] = H(X)$

Consequences of the AEP: the typical set

Definition: $A_\epsilon^{(n)}$ is a typical set with respect to $P_X(x)$ if it is the set of sequences in the set of all possible sequences $x_1^n \in \underline{\mathcal{X}}^n$ with probability:

$$2^{-n(H(X)+\epsilon)} \leq P\left(X_1^n = x_1^n\right) \leq 2^{-n(H(X)-\epsilon)}$$

equivalently

$$H(X) - \epsilon \leq -\frac{1}{n}\log(P\left(X_1^n = x_1^n\right)) \leq H(X) - \epsilon$$

The bounds can be made arbitrarily tight as n increases.

Consequences of the AEP: the typical set

Typical:

$$P(X_1^n \in A_\epsilon^{(n)}) \to 1$$

Notice: two different limits, $\forall \epsilon, \delta > 0$, $\exists N$, s.t. $n \geq N$ implies

$$P(A_\epsilon^{(n)}) \geq 1 - \delta$$

For simplicity, set $\delta = \epsilon$.

How big is the typical set?

Size of the Typical Set

Claim:

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

and for large enough n,

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

Proof:

$$\begin{aligned} 1 &= \sum_{\mathcal{X}^n} P(x_1^n) \\ &\geq \sum_{A_\epsilon^{(n)}} P(x_1^n) \\ &\geq |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)} \end{aligned}$$

For large enough n,

$$\begin{aligned} 1-\epsilon &\leq P(A_\epsilon^{(n)}) \\ &= \sum_{A_\epsilon^{(n)}} P(x_1^n) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)} \end{aligned}$$

Compare to $|\mathcal{X}^n| = 2^{n \log |\mathcal{X}|}$.

Example

Consider binary r.v.s X_i, i.i.d. with $P(X = 0) = p$, and $P(X = 1) = 1 - p$.

A "typical" sequence of length n has roughly np 0's and $n(1-p)$ 1's, the probability for that to happen is

$$\begin{aligned} p^{np}(1-p)^{n(1-p)} &= 2^{n(p\log p + (1-p)\log(1-p))} \\ &= 2^{-nH(X)} \end{aligned}$$

How many "typical" sequences are there?

Stirling Formula $n! \approx n^n e^{-n}\sqrt{2\pi n}$.

$$\begin{aligned} \binom{n}{np} &= \frac{n!}{(np)!(n(1-p))!} \\ &\approx \frac{n^n e^{-n}}{(np)^{np} e^{-np} (n(1-p))^{n(1-p)} e^{-n(1-p)}} \\ &= \frac{1}{p^{np}(1-p)^{n(1-p)}} \\ &= 2^{nH(X)} \end{aligned}$$

What about the ϵ?

H is continuous in p.

Let $p < 1/2$, what about I take the set of the most likely sequences, i.e., those with less than np 0's?

Notation: $H(p) = -p \log p - (1-p)\log(1-p)$.

$$\begin{aligned} & \sum_{t: nt \in \mathcal{Z}, t \leq p} \binom{n}{nt} \\ \approx & \sum_{t: nt \in \mathcal{Z}, t \leq p} 2^{nH(t)} \\ \approx & \ 2^{nH(p)} \end{aligned}$$

It doesn't change the size too much.

Using the Typical Set for Data Compression

Description in typical set requires no more than $n(H(X) + \epsilon) + 1$ bits (correction of 1 bit because of integrality)

Description in atypical set ${A_\epsilon^{(n)}}^C$ requires no more than $n \log(\mathcal{X}) + 1$ bits

Add another bit to indicate whether in $A_\epsilon^{(n)}$ or not to get whole description

Consequences of the AEP: using the typical set for compression

Let $l(x_1^n)$ be the length of the binary description of x_1^n

$\forall \epsilon > 0$, $\exists n_0$ s.t. $\forall n > n_0$,

$$\begin{aligned}
& E_{X_1^n}[l(X_1^n)] \\
= & \sum_{x_1^n \in A_\delta^{(n)}} P_{X_1^n}(x_1^n)\, l(x_1^n) + \sum_{x_1^n \in {A_\delta^{(n)}}^C} P_{X_1^n}(x_1^n)\, l(x_1^n) \\
\leq & \sum_{x_1^n \in A_\delta^{(n)}} P_{X_1^n}(x_1^n)\, (n(H(X)+\delta)+2) \\
+ & \sum_{x_1^n \in {A_\delta^{(n)}}^C} P_{X_1^n}(x_1^n)\, (n \log(|\mathcal{X}|)+2) \\
= & nH(X) + n\epsilon
\end{aligned}$$

for δ small enough with respect to ϵ

so $E_{X_1^n}[\frac{1}{n} l(X_1^n)] \leq H(X) + \epsilon$ for n sufficiently large.