LECTURE 3

Convergence and Asymptotic Equipartition Property

Last time:

- Convexity and concavity
- Jensen's inequality
- Positivity of mutual information
- Data processing theorem
- Fano's inequality

Lecture outline

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

Reading: Scts. 3.1-3.2.

Types of convergence

Recall what a random variable is: a mapping from its set of sample values Ω onto $\mathcal{R}$

$$\begin{aligned} X : \quad & \Omega \mapsto \mathcal{R} \\ & \xi \rightarrow X(\xi) \end{aligned}$$

In the cases we have been discussing, $\Omega = \mathcal{X}$ and we map onto $[0, 1]$

Types of convergence

- Sure convergence: a random sequence $X_1, \ldots$ converges surely to r.v. X if $\forall \xi \in \Omega$ the sequence $X_n(\xi)$ converges to $X(\xi)$ as $n \to \infty$
- Almost sure convergence (also called convergence with probability 1) the random sequence converges a.s. (w.p. 1) to X if the sequence $X_1(\xi), \ldots$ converges to $X(\xi)$ for all ξ except possibly on a set of Ω of probability 0
- Mean-square convergence: $X_1, \ldots$ converges in m.s. sense to r.v. X if

$$\lim_{n\to\infty} E_{X_n}[|X_n - X|^2] \to 0$$

- Convergence in probability: the sequence converges in probability to X if $\forall \epsilon > 0$

$$\lim_{n\to\infty} Pr[|X_n - X| > \epsilon] \to 0$$

- Convergence in distribution: the sequence converges in distribution if the cumulative distribution function $F_n(x) = Pr(X_n \leq x)$ satisfies $\lim_{n\to\infty} F_n(x) \to F_X(x)$ at all x 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

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

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

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

$$\text{Pr}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma_X^2}{n\epsilon^2}$$

Weak Law of Large Numbers

Consequence of Chebyshev's inequality: Random variable X

$$\sigma_X^2 = \sum_{x \in \mathcal{X}} (x - \mathbf{E}[X])^2 P_X(x)$$

$$\sigma_X^2 \;\geq\; c^2 \, \mathrm{Pr}(|X - \mathbf{E}[X]| \geq c)$$

$$\mathrm{Pr}(|X - \mathbf{E}[X]| \geq c) \;\leq\; \frac{\sigma_X^2}{c^2}$$

$$\mathrm{Pr}(|X - \mathbf{E}[X]| \geq k\sigma_X) \;\leq\; \frac{1}{k^2}$$

Strong Law of Large Numbers

Theorem: (SLLN) If X_i are IID, and $E_X[|X|] < \infty$, then

$$M_n = \frac{X_1 + \cdots + X_n}{n} \to E_X[X], \qquad \text{w.p.1.}$$

AEP

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

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

Notation: $\underline{X}_i^j = (X_i, \ldots, X_j)$ (if $i = 1$, generally omitted)

Proof: create r.v. Y that 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

$$
\begin{aligned}
& -\frac{1}{n}\log(P_{\underline{X}^n}(\underline{x}^n)) \\
= \; & -\frac{1}{n}\sum_{i=1}^{n}\log(P_X(x_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}
$$

using the WLLN on Y

$\frac{1}{n}\sum_{i=1}^{n} y_i \rightarrow E_Y[Y]$ in probability

$E_Y[Y] = -E_Z[\log(P_X(Z))] = H(X)$

for some r.v. Z identically distributed with 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 $\underline{x}^n \in \underline{\mathcal{X}}^n$ with probability:

$$2^{-n(H(X)+\epsilon)} \leq P_{\underline{X}^n}(\underline{x}^n) \leq 2^{-n(H(X)-\epsilon)}$$

equivalently

$$H(X) - \epsilon \leq -\frac{1}{n}\log(P_{\underline{X}^n}(\underline{x}^n)) \leq H(X) + \epsilon$$

As n increases, the bounds get closer together, so we are considering a smaller range of probabilities

We shall use the typical set to describe a set with characteristics that belong to the majority of elements in that set.

Note: the variance of the entropy is finite

Consequences of the AEP: the typical set

Why is it typical? AEP says $\forall \epsilon > 0$, $\forall \delta > 0$, $\exists n_0$ such that $\forall n > n_0$

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

(note: δ can be ϵ)

How big is the typical set?

$$\begin{aligned}
1 &= \sum_{\underline{x}^n \in \mathcal{X}^n} P_{\underline{X}^n}(\underline{x}^n) \\
&\geq \sum_{\underline{x}^n \in A_\epsilon^{(n)}} P_{\underline{X}^n}(\underline{x}^n) \\
&\geq \sum_{\underline{x}^n \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\
&= |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)} \\
&\Rightarrow |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}
\end{aligned}$$

$$
\begin{aligned}
& Pr(A_\epsilon^{(n)}) \geq (1-\epsilon) \\
\Rightarrow \quad & 1-\epsilon \leq \sum_{\underline{x}^n \in A_\epsilon^{(n)}} P_{\underline{X}^n}(\underline{x}^n) \\
\leq \quad & |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)} \\
\Rightarrow \quad & |A_\epsilon^{(n)}| \geq 2^{n(H(X)-\epsilon)}(1-\epsilon)
\end{aligned}
$$

Visualize:

Consequences of the AEP: using the typical set for 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 $l(\underline{x}^n)$ be the length of the binary description of $\underline{x}^n$

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

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

for δ small enough with respect to ϵ

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

Jointly typical sequences

$A_\epsilon^{(n)}$ is a typical set with respect to $P_{X,Y}(x,y)$ if it is the set of sequences in the set of all possible sequences $(\underline{x}^n, \underline{y}^n) \in \underline{\mathcal{X}}^n \times \underline{\mathcal{Y}}^n$ with probability:

$$2^{-n(H(X)+\epsilon)} \leq P_{\underline{X}^n}(\underline{x}^n) \leq 2^{-n(H(X)-\epsilon)}$$

$$2^{-n(H(Y)+\epsilon)} \leq P_{\underline{Y}^n}\left(\underline{y}^n\right) \leq 2^{-n(H(Y)-\epsilon)}$$

$$2^{-n(H(X,Y)+\epsilon)} \leq P_{\underline{X}^n,\underline{Y}^n}\left(\underline{x}^n, \underline{y}^n\right) \leq 2^{-n(H(X,Y)-\epsilon)}$$

for $(\underline{X}^n, \underline{Y}^n)$ sequences of length n IID according $P_{\underline{X}^n,\underline{Y}^n}(\underline{x}^n, \underline{y}^n) = \prod_{i=1}^n P_{X,Y}(x_i, y_i)$

$Pr((\underline{X}^n, \underline{Y}^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$

Jointly typical sequences

Use the union bound

$$
\begin{aligned}
& Pr((\underline{X}^n, \underline{Y}^n) \notin A_\epsilon^{(n)}) \\
\leq \; & Pr((\underline{X}^n, \underline{Y}^n) \notin A_\epsilon^{\prime\prime\prime(n)}) \\
+ \; & Pr((\underline{X}^n) \notin A_\epsilon^{\prime\prime(n)}) \\
+ \; & Pr((\underline{Y}^n) \notin A_\epsilon^{\prime(n)})
\end{aligned}
$$

For A''' single typical sequence for pair, A'' for X and A' for Y

each element in the RHS goes to 0