Information Theory: Lecture Notes 3

zqy1018

June 5, 2020

Contents

1 Asymptotic Equipartition Property

The i.i.d. random variables have LLN in probability theory. In information theory, they have AEP in turn.

Theorem 1. (Asymptotic Equipartition Property, AEP) If $X_1, X_2, \cdots$ are i.i.d. then

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) \xrightarrow{p} H(X)$$

Proof. By weak LLN,

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) = -\frac{1}{n}\sum_i \log p(X_i) \xrightarrow{p} -\mathrm{E}\log p(X) = H(X)$$

□

Note. We can write AEP theorem in epsilon-N language:

$$\forall \epsilon > 0, \forall \delta > 0, \exists N \in \mathbb{N}, \forall n > N, p\left(\left|-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) - H(X)\right| < \epsilon\right) > 1 - \delta$$

2 Typicality

The typical set is the set of typical sequences with almost the same probability.

Definition. Assuming $X_1, X_2, \cdots$ are i.i.d., the **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of **typical sequences** $(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

Note. Sometimes we use x^n to denote the sequence $(x_1, x_2, \cdots, x_n)$.

Theorem 2.

(1) $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$ iff $H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) \leq H(X) + \epsilon$.
(2) $p(A_\epsilon^{(n)}) \geq 1 - \epsilon$ for n sufficiently large.
(3) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.
(4) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Proof.

For (1), by definition.

For (2), by the epsilon-N form of AEP, letting $\delta = \epsilon$.

For (3), we use a classical proof methodology in probability theory. We have

$$1 = \sum_{x\in\mathcal{X}^n} p(x) \geq \sum_{x\in A_\epsilon^{(n)}} p(x) \geq |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)} \implies |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$$

For (4), using (2):

$$1-\epsilon < p(A_\epsilon^{(n)}) \leq \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = |A_\epsilon^{(n)}|2^{-n(H(X)-\epsilon)} \implies |A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$$

□

Note. These 4 theorems all show some important properties about typical sequences.

(1) can be used to judge whether a sequence is typical or not. And it shows that all elements of the typical set have nearly the same probability.

(2) shows that the typical set has probability nearly 1 when n is sufficiently large.

(3) and (4) shows that the number of elements in the typical set is nearly 2^{nH} when n is sufficiently large. And (3) shows the *upper bound* of the size of a typical set.

(3) also shows that the typical set can be much smaller than the original set, since $\frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} \leq 2^{n(H(X)-\log|\mathcal{X}|)}$. If X does not obey a uniform distribution, then $2^{n(H(X)-\log|\mathcal{X}|)} \to 0$ when $n \to \infty$.

3 High Probability Set

We can see that $A_\epsilon^{(n)}$ is a tiny set with high probability. Is it the smallest?

Theorem 3. Assuming $X_1, X_2, \cdots$ are i.i.d.. Let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the *smallest* set with $p(B_\delta^{(n)}) \geq 1-\delta$. Then for $\delta < \frac{1}{2}$ and any $\delta' > 0$, for n sufficiently large,

$$\frac{1}{n}\log|B_\delta^{(n)}| > H(X) - \delta'$$

Proof. Since $p(B_\delta^{(n)}) \geq 1-\delta, p(A_\epsilon^{(n)}) \geq 1-\epsilon$, we have

$$1-\epsilon-\delta \leq p(A_\epsilon^{(n)} \cap B_\delta^{(n)}) = \sum_{x\in A_\epsilon^{(n)}\cap B_\delta^{(n)}} p(x) \leq |A_\epsilon^{(n)} \cap B_\delta^{(n)}|2^{-n(H(X)-\epsilon)} \leq |B_\delta^{(n)}|2^{-n(H(X)-\epsilon)}$$

So $|B_\delta^{(n)}| \geq (1-\epsilon-\delta)2^{n(H(X)-\epsilon)}$. Hence we can choose a good ϵ and n to achieve the bound. □

By this theorem, $B_\delta^{(n)}$ must have at least $2^{nH(X)}$ elements, to first order in the exponent. But $A_\epsilon^{(n)}$ has $2^{n(H(X)\pm\epsilon)}$ elements. Therefore, $A_\epsilon^{(n)}$ is about the same size as the smallest high-probability set.

We will now define some new notation to express equality to first order in the exponent.

Definition. If $\lim_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} = 0$, then a_n and b_n are **equal to the first order in the exponent**. Use $a_n \doteq b_n$ to denote it.

Corollary. If $\delta_n \to 0$ and $\epsilon_n \to 0$, then

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}$$

4 Joint Typicality

For two random variables, we can define the joint typicality between them.

Definition. the **jointly typical set** $A_\epsilon^{(n)}$ with respect to $p(x^n, y^n)$ is the set of **jointly typical sequences** $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ with the property

$$\begin{aligned} A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ \left| -\frac{1}{n} \log p\left(x^n\right) - H(X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p\left(y^n\right) - H(Y) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p\left(x^n, y^n\right) - H(X,Y) \right| < \epsilon \Big\} \end{aligned}$$

where

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$$

Warning. $x^n \in A_{x,\epsilon}^{(n)} \wedge y^n \in A_{y,\epsilon}^{(n)}$ does not imply $(x^n, y^n) \in A_\epsilon^{(n)}$.

For the jointly typical set, there are some similar properties.

Theorem 4.

(1) $p(A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.

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

(3) If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, i.e. $\tilde{X}^n, \tilde{Y}^n$ are independent with the same marginal distributions as $p(x^n, y^n)$, then

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \leq 2^{-n(I(X;Y)-3\epsilon)}$$

and for sufficiently large n,

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Proof. We only show the proof for (3) here. For (1), the proof uses three conditions in the definition. For (2), the proof is almost the same as before.

For (3), by (2), we have

$$\begin{aligned} \Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\ &= 2^{-n(I(X;Y)-3\epsilon)} \end{aligned}$$

And by (1), for sufficiently large n, we have

$$1-\epsilon < p(A_\epsilon^{(n)}) \leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n,y^n) \leq |A_\epsilon^{(n)}|2^{-n(H(X,Y)-\epsilon)} \implies |A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$$

So

$$\begin{aligned}\Pr\left\{(\tilde{X}^n,\tilde{Y}^n)\in A_\epsilon^{(n)}\right\} &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n)\\ &\geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}\\ &= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}\end{aligned}$$

□

Note. We can see the size of the jointly typical set is about $2^{nH(X,Y)}$. And if we randomly pick a typical sequence x^n from the typical set of $\mathcal{X}^n$, and randomly pick another y^n from the typical set of $\mathcal{Y}^n$, then

$$\Pr\left\{(x^n,y^n)\in A_\epsilon^{(n)}\right\} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}$$

5 Review

The typical set shows an interesting phenomenon in the set of i.i.d. sequences. And we will use the joint typicality later in the proof of channel coding theorem.

Acknowledgment

The contents are mainly based on the course materials of CS258, 2020 Spring, Shanghai Jiao Tong University and *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas.