EL-GY 6063: Information Theory
Lecture 6

March 2, 2018

1 The Asymptotic Equipartition Property (AEP)

1.1 Weak Typicality for Discrete Random Variables

Review:

Theorem 1 (AEP). *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables. Then,*

$$P(|-\frac{1}{n}\log P_{X^n}(X^n) - H(X)| > \epsilon) \to 0 \text{ as } n \to \infty,$$

For any $\epsilon > 0$.

Definition 1. *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables. The weakly typical set of sequences of length n with respect to P_X is defined as:*

$$\mathcal{A}_\epsilon^n(X) = \{x^n || -\frac{1}{n}\log P_{X^n}(x^n) - H(X)| \leq \epsilon\}.$$

The sequence x^n is called an ϵ-(weakly) typical sequence with respect to P_X if $x^n \in \mathcal{A}_\epsilon^n(X)$.

Theorem 2. *The set $\mathcal{A}_\epsilon^n(X)$ satisfies the following properties:*
i) $P(\mathcal{A}_\epsilon^n(X)) \to 1$ as $n \to \infty$. Alternatively, for any $\epsilon > 0$ and for sufficiently large n, we have $P(\mathcal{A}_\epsilon^n(X)) > 1 - \epsilon$,
ii) $|\mathcal{A}_\epsilon^n(X)| \leq 2^{n(H(X)+\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^n(X)|$.

Definition 2. *Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables. The weakly typical set of sequences of length n with respect to $P_{X,Y}$ is defined as:*

$$\mathcal{A}_\epsilon^n(X,Y) = \{(x^n, y^n) \in \mathcal{A}_\epsilon^n(X) \times \mathcal{A}_\epsilon^n(Y) \big|| -\frac{1}{n}\log P_{X^n,Y^n}(x^n, y^n) - H(X,Y)| \leq \epsilon\}.$$

The sequence pair (x^n, y^n) is called an ϵ-(weakly) typical sequence with respect to $P_{X,Y}$ if $(x^n, y^n) \in \mathcal{A}_\epsilon^n(X,Y)$.

Theorem 3. *The set $\mathcal{A}_\epsilon^n(X,Y)$ satisfies the following properties:*
i) $P(\mathcal{A}_\epsilon^n(X,Y)) \to 1$ as $n \to \infty$. Alternatively, for any $\epsilon > 0$ and for sufficiently large n, we have $P(\mathcal{A}_\epsilon^n(X,Y)) > 1 - \epsilon$,
ii) $|\mathcal{A}_\epsilon^n(X,Y)| \leq 2^{n(H(X,Y)+\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |\mathcal{A}_\epsilon^n(X,Y)|$.

The conditional typical set is defined as follows:

Definition 3. *Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables. The set of conditionally weakly typcial sequences of length n with respect to the sequence y^n and distribution $P_{X|Y}$ is defined as:*

$$\mathcal{A}_\epsilon^n(X|Y^n = y^n) = \{x^n | (x^n, y^n) \in \mathcal{A}_\epsilon^n(X,Y)\}.$$

Theorem 4. *If $y^n \notin \mathcal{A}_\epsilon^n(Y)$, then $\mathcal{A}_\epsilon^n(X|Y^n = y^n) = \phi$. Otherwise If $y^n \in \mathcal{A}_\epsilon^n(Y)$, then,*
i) $P(\mathcal{A}_\epsilon^n(X|Y^n = y^n)|Y^n = y^n) \to 1$ as $n \to \infty$.
ii) $|\mathcal{A}_\epsilon^n(X|Y^n = y^n)| \leq 2^{n(H(X|Y)+2\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X|Y)-2\epsilon)} \leq |\mathcal{A}_\epsilon^n(X|Y^n = y^n)|$.

1.2 Types of Sequences and Strong Typicality

Strong typicality is sometimes also called frequency typicality. It related to the frequencies of symbols. We will briefly touch upon the notion of strong typicality. For a more complete discussion refer to Ch.1 of Csiszar and Korner's book.

Theorem 5. *Let X^n be a sequence of independent and identically distributed random variables distributed according to P_X. Let $N(a|x^n), a \in \mathcal{X}$ be the number of indices $i \in [1, n]$ such that $x_i = a$. Then, $P(\exists a \in \mathcal{X} : |\frac{1}{n}N(a|X^n) - P_X(a)| \geq \epsilon) \to 1$ as $n \to \infty$.*

proof.

$$P(\exists a \in \mathcal{X} : |\frac{1}{n}N(a|X^n) - P_X(a)| \geq \epsilon) = P(\cup_{a\in\mathcal{X}}|\frac{1}{n}N(a|X^n) - P_X(a)| \geq \epsilon) \leq \sum_{a\in\mathcal{X}} P(|\frac{1}{n}N(a|X^n) - P_X(a)| \geq \epsilon).$$

Each of the terms in the last summation go to 0 by the weak law of large numbers. From calculus, we know that the finite addition of sequences each of which approaches 0 as $n \to \infty$ goes to 0 as well.

Definition 4. *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables. The strongly typical set of sequences of length n with respect to P_X is defined as:*

$$\mathcal{T}_\epsilon^n(X) = \{x^n | \forall a \in \mathcal{X} : |\frac{1}{n}N(a|X^n) - P_X(a)| \leq \epsilon\}.$$

The sequence x^n is called an ϵ-(strongly) typical sequence with respect to P_X if $x^n \in \mathcal{T}_\epsilon^n(X)$.

Theorem 6. *The set $\mathcal{T}_\epsilon^n(X)$ satisfies the following properties:*
i) $P(\mathcal{T}_\epsilon^n(X)) \to 1$ as $n \to \infty$.
ii) $|\mathcal{T}_\epsilon^n(X)| \leq 2^{n(H(X)+f(\epsilon))}$,
iii)$(1-\epsilon)2^{n(H(X)-f(\epsilon))} \leq |\mathcal{T}_\epsilon^n(X)|$, where $f(\cdot)$ is a continuously increasing function for which $f(0) = 0$.

Joint typicality and conditional typicality defined similarly.

Example: i) Consider the set of all binary sequences of length n: $\{x^n | x_i \in \{0, 1\}\}$. Let $k \in \{1, 2, \cdots, n\}$. Count the number of binary sequences x^n with k ones and $n - k$ zeros.
Solution. There are $\binom{n}{k}$ such sequences. These are called sequences of type $(\frac{k}{n}, \frac{n-k}{n})$.

Define the set of all such sequences $\mathcal{T}_{\frac{k}{n}}^n = \{x^n | x^n$ has k ones$\}$. So far, we have shown that $|\mathcal{T}_{\frac{k}{n}}^n| = \binom{n}{k}$.

This result can be simplified further.

Lemma 1 (Stirling's approximation)**.** *For any natural number n, we have:*

$$\sqrt{2\pi}(n^{n+\frac{1}{2}})e^{-n} \leq n! \leq e(n^{n+\frac{1}{2}})e^{-n}.$$

Lemma 2. *For any $\delta > 0$, we have:*

$$2^{n(H(\frac{k}{n})-\delta)} \leq \binom{n}{k} \leq 2^{n(H(\frac{k}{n})+\delta)}.$$

proof.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{(n^{n+\frac{1}{2}})e^{-n}}{(k^{k+\frac{1}{2}})e^{-k}((n-k)^{(n-k)+\frac{1}{2}})e^{-n-k}} = \sqrt{\frac{n}{k(n-k)}}(\frac{k}{n})^{-k}(\frac{n-k}{n})^{-(n-k)}$$
$$= \sqrt{\frac{n}{k(n-k)}}2^{n(-\frac{k}{n}\log\frac{k}{n} - \frac{n-k}{n}\log\frac{n-k}{n})} = \sqrt{\frac{n}{k(n-k)}}2^{-nH(\frac{k}{n})}.$$

The proof is completed by noting that $2^{-n\delta} \leq \sqrt{\frac{n}{k(n-k)}} \leq 2^{n\delta}$ for sufficiently large n.

ii) Let $X_i, i = 1, 2, \cdots$ be independent and identically distributed Bernoulli random variables with parameter p. Let $x^n \in \mathcal{T}^n_{\frac{k}{n}}$. Calculate $P_{X^n}(x^n)$.
Solution. $P_{X^n}(x^n) = p^k(1-p)^{n-k}$.

Note that the probability is the same for all sequences in the same type. This can be simplified as follows:

$$P_{X^n}(x^n) = p^k(1-p)^{n-k} = 2^{\log p^k} 2^{(1-p)^{n-k}} = 2^{k \log p + (n-k) \log 1 - p} = 2^{n(\frac{k}{n} \log p + \frac{n-k}{n} \log 1 - p)},$$

let $q = \frac{k}{n}$. Then,

$$P_{X^n}(x^n) = 2^{n(q \log p + (1-q) \log 1 - p)} = 2^{-n(D(Q_X||P_X) + H(Q_X))}.$$

iii) Calculate $P(\mathcal{T}^n_{\frac{k}{n}})$ in terms of P_X and Q_X.
Solution. $P(\mathcal{T}^n_{\frac{k}{n}}) = \sum_{x^n \in \mathcal{T}^n_{\frac{k}{n}}} P_{X^n}(x^n) = \sum_{x^n \in \mathcal{T}^n_{\frac{k}{n}}} p^k(1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$.

This is a familiar result. The probability distribution is a binomial with parameters (n, p). Note that from the above, we can conclude that:

$$\begin{aligned} &2^{n(H(Q_X)-\delta)} 2^{-n(D(Q_X||P_X)+H(Q_X))} \leq P(\mathcal{T}^n_{\frac{k}{n}}) \leq 2^{n(H(Q_X)+\delta)} 2^{-n(D(Q_X||P_X)+H(Q_X))} \\ &\Rightarrow 2^{-n(D(Q_X||P_X)+\delta)} \leq P(\mathcal{T}^n_{\frac{k}{n}}) \leq 2^{-n(D(Q_X||P_X)-\delta)}, \end{aligned}$$

for any $\delta > 0$, where Q_X is the distribution of the Bernoulli random variable with parameter $\frac{k}{n}$ and P_X is the distribution of the Bernoulli random variable with parameter p.

2 Continuous alphabet Random Variables and Continuous Random Variables

Reference: Chapter 8 of Elements of Information Theory.

Throughout this course whenever we talk about continuous alphabet random variables we are referring to random variables with alphabet $\mathcal{X} = \mathbb{R}$.

Reminder:

Definition 5. *Discrete Random Variable: A discrete random variable X is characterized by $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$, where $\mathcal{X} \subset \mathbb{R}$ is finite.*

Generally, any random variable is characterized by three components. The sample space, the event space and the probability measure. In this course the sample space for continuous random variables is taken to be the real line. The event space is taken to be the Borel sigma field which is defined below:

Definition 6. *The Borel sigma field $\mathbb{B}$ is defined as the smallest sigma field containing all closed intervals in the real line. Alternatively, it is the set of all countable unions, intersections and complements of sets of the form $[a, b]$.*

Example. The Borel sigma field determines the set of events which can be defined for the experiment. For example, Assume that the random variable X is distributed uniformly on the unit interval. In this case, $\mathcal{E} = [0, 0.5]$ is an event in the Borel sigma field. Also, its complement $\mathcal{E}^c = (-\infty, 0) \cup (0.5, \infty)$ is another event. $P(\mathcal{E}) = 0.5$ and $P(\mathcal{E}^c) = 0.5$.

Example. The set with a single element $\{0\}$ is in the Borel Sigma field since it can be written as $\cap_{i=1}^{\infty}[0, \frac{1}{n}]$.
Example. The set Q^c where Q is the set of rational numbers is in the Borel sigma field since Q is a countable union of elements of the Borel Sigma field.

The probability measure of a random variable is a function $P_X : \mathbb{B} \to [0, 1]$ which satisfies the axioms of probability.

Note that the event $\mathcal{E}_a = (-\infty, a), a \in \mathbb{R}$ is an element of the Borel sigma field. So, $P_X((-\infty, a))$ is well-defined. This is called the cumulative distribution function or cdf of the random variable.

Definition 7. *For the random variable X characterized by $(\mathbb{R}, \mathbb{B}, P_X)$, the function $F_X(a) = P_X((-\infty, a)) = P(X \leq a)$ is called the cdf of X.*

It follows from the axioms of probability that the cdf $F_X(a)$ is a non-decreasing function.

proof. Let $a < b$, then $(-\infty, a) \subset (-\infty, b)$ so $P_X((-\infty, a)) \leq P_X((-\infty, b))$ which itself follow from Axiom 3 and 1.

In this class, we are only interested in random variables for which $F_X(\cdot)$ is a differentiable function. These are called continuous random variables. For these variables the probability density function or pdf is defined as the derivative of the cdf.

Definition 8. *For the continuous random variable X characterized by $(\mathbb{R}, \mathbb{B}, P_X)$, the function $f_X(a) = \frac{d}{dx}F_X(x)|_{x=a}$ is called the pdf of X.*

Example. Assume that X is a uniform random variable on the unit interval. In this case:

$$F_X(a) = P_X((-\infty, a)) = \begin{cases} 0 & \text{if } a \leq 0 \\ a & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } 1 \leq a \end{cases} .$$

So,

$$f_X(a) = \frac{d}{dx}F_X(x)|_{x=a} = \begin{cases} 0 & \text{if } a \leq 0 \text{ or } 1 \leq a \\ 1 & \text{if } 0 \leq a \leq 1 \end{cases} .$$

Question. Let X be uniform on the unit interval. What is the probability of the midpoint $P_X(\{0.5\})$?

Note that for a continuous random variable, the cdf $F_X(a)$ is continuous since it is differentiable. As a result, $P_X(\{a\}) = \lim_{\epsilon \to 0} F_X(a + \epsilon) - F_X(a) = 0, \forall b \in \mathbb{R}$. So, every point has 0 probability for continuous random variables.

Similarly, one can define a pair of continuous random variables.

Definition 9. *For the pair of random variables (X, Y) characterized by $(\mathbb{R}^2, \mathbb{B}^2, P_{X,Y})$, the function $F_{X,Y}(a, b) = P_{X,Y}((-\infty, a) \times (-\infty, b))$ is called the joint cdf of (X, Y).*

Definition 10. *For the pair of random variables (X, Y) characterized by $(\mathbb{R}^2, \mathbb{B}^2, P_{X,Y})$, the function $f_{X,Y}(a, b) = \frac{d^2}{dadb}F_{X,Y}(a, b)$ is called the joint pdf of (X, Y).*

The conditional pdf is defined as follows:

Definition 11. *For the pair of random variables (X, Y) characterized by $(\mathbb{R}^2, \mathbb{B}^2, P_{X,Y})$, and the joint pdf $f_{X,Y}(a, b)$, the conditional pdf of X given Y is defined as $f_{X|Y}(a|b) = \frac{f_{XY}(a,b)}{f_Y(b)}$.*

2.1 Measurement of Resolution Δ

How much uncertainty is there in a continuous random variable? In the lossless source coding chapter, we saw that there is an alternative way of asking this question, that is, how many bits does it take to compress the random variable. As an example, let us take the random variable X which is uniform on the unit interval. How many decimal places does it take to write X? The number is not limited, it takes a possibly infinite number of decimal places. How many bits does it take to losslessly represent X? It follows that it takes a possibly infinite number of bits to represent X. That is, it seems that the amount of uncertainty in X is infinite bits. We show this mathematically in the next part of the lecture.

Let us consider an arbitrary random variable X for which the pdf f_X is continuous. Fix $\Delta > 0$. By the mean value theorem, for any interval $[i\Delta, (i + 1)\Delta]$, there exists x_i in the interval, such that:

$$\int_{x=i\Delta}^{(i+1)\Delta} f(x)dx = f_X(x_i)\Delta.$$

Consider the quantization of X which is defined by $Y = Q(X) = x_i$ if $X \in [i\Delta, (i+1)\Delta]$.
Then Y is a discrete random variable with:

$$P_Y(x_i) = P(X \in [i\Delta, (i+1)\Delta]) = \int_{x=i\Delta}^{(i+1)\Delta} f(x)dx = f_X(x_i)\Delta.$$

So,

$$H(X) \geq H(Y) = -\sum_y P_Y(y) \log P_Y(y) = -\sum_i P_X(x_i) \log P_X(x_i) = -\sum_i f_X(x_i)\Delta \log f_X(x_i)\Delta$$
$$= -\sum_i f_X(x_i)\Delta \log f_X(x_i) - \log \Delta = -\int_x f_X(x) \log f_X(x)dx - \log \Delta.$$

Which goes to ∞ as Δ becomes smaller.

Example. Let X be uniform on the unit interval and $\Delta = \frac{1}{8}$. Then Y is uniform on alphabet 8, so $H(Y) = 3$ which can also be seen in the above formula. Also, Let $\Delta = 2^{-n}$. Then, $H(Y) = n$. In other words, at least n bits of info is necessary to describe Y which is the quantized version of X.

2.2 Differential Entropy

Definition 12. *For the continuous random variable X characterized by $(\mathbb{R}, \mathbb{B}, P_X)$, and the pdf $f_X(a)$ the differential entropy of X is defined as:*

$$h_d(X) = -\int_{x\in\mathbb{R}} f_X(x) \log f_X(x)dx.$$

Remember that for discrete variables we sometimes wrote $H(P_X)$ instead of $H(X)$. For continuous variables, we sometimes write $h_d(f_X)$ instead of $h_d(X)$.

As mentioned above, the differential entropy does not have the same operational meaning as entropy and it is often very counter intuitive. One must avoid interpreting differential entropy as a measure of the uncertainty in a source.

Example. Let X be the uniform random variable over the unit interval. Calculate $h_d(X)$.

Solution. $h_d(X) = -\int_{x\in\mathbb{R}} f_X(x) \log f_X(x)dx = -\int_{x\in[0,1]} 1 \cdot \log 1dx = 0$. As we stated before, the uncertainty in the random variable X is infinity but the differential entropy is 0.

Example. Let X be uniformly distributed over the interval $[a, b]$. Calculate $h_d(X)$.

Solution. $h_d(X) = -\int_{x\in\mathbb{R}} f_X(x) \log f_X(x)dx = -\int_{x\in[a,b]} \frac{1}{a-b} \cdot \log \frac{1}{a-b}dx = \log a - b$. Let $a = 0$ and $b = 0.5$. Then, $h_d(X) = -1$. So, the differential entropy may be negative.

Example. Let X be uniformly distributed over the unit interval. Let $Y = 2X$. Compare $h_d(Y)$ and $h_d(X)$.

Solution. Clearly, Y is uniform over $[0, 2]$. So, $h_d(Y) = 1$ and $h_d(X) = 0$. So, although multiplication by 2 is a one-to-one transformation, the differential entropy is increased.

Example. Let X be a continuous random variable and let $Y = aX$. Write $h_d(Y)$ in terms of $h_d(X)$ and $a > 0$.

Solution. Note that $F_Y(y) = P(Y \leq y) = P(aX \leq y) = P(X \leq \frac{y}{a}) = F_X(\frac{y}{a})$. So, $f_Y(y) = \frac{1}{a}f_X(\frac{y}{a})$. So,

$$h_d(Y) = -\int_{y\in\mathbb{R}} f_Y(y) \log f_Y(y)dy = -\int_{y\in\mathbb{R}} \frac{1}{a}f_X(\frac{y}{a}) \log \frac{1}{a}f_X(\frac{y}{a})dy = -\int_{x\in\mathbb{R}} f_X(x) \log \frac{1}{a}f_X(x)dx = h_d(X) + \log a.$$

Example. Let $X \sim N(\eta, \sigma)$ calculate $h_d(X)$.

Solution. See Example 8.2.1 in the textbook.

Definition 13. *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed continuous random variables. The weakly typical set of sequences of length n with respect to f_X is defined as:*

$$\mathcal{A}_\epsilon^n(X) = \{x^n \big| | -\frac{1}{n} \log f_{X^n}(x^n) - h_d(X)| \leq \epsilon\}.$$

The sequence x^n is called an ϵ-(weakly) typical sequence with respect to f_X if $x^n \in \mathcal{A}_\epsilon^n(X)$.

Definition 14. *The volume $Vol(\mathcal{A})$ of a set $\mathcal{A} \in \mathbb{R}^n$ is defined as*

$$Vol(\mathcal{A}) = \int_{\mathcal{A}} dx^n.$$

Theorem 7. *The weakly typical set $\mathcal{A}_\epsilon^n(X)$ for the continuous variable X satisfies:*
i) $P(\mathcal{A}_\epsilon^n(X) \to 1$ as $n \to \infty$.
ii) $Vol(\mathcal{A}_\epsilon^n(X)) \leq 2^{n(H(X)+\epsilon)}$.
iii) $(1-\epsilon)2^{n(H(X)-\epsilon)} \leq Vol(\mathcal{A}_\epsilon^n(X))$ for n sufficiently large.