ECE253/CSE208 Introduction to Information Theory

Lecture 5: Asymptotic Equipartition Property (AEP)

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 3 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*
- Lecture 1 (typical sequences) of *Information Theory for Wireless Comms.* by Dr. Saif K. Mohammed

Law of Large Numbers (LLN) in Information Theory

- LLN in probability — Sample mean converges to the true mean: $\bar{X} = \frac{1}{n}\sum_i X_i \xrightarrow[n\to\infty]{\text{i.p.}} \mathrm{E}(X)$.
- LLN in info theory — Sample entropy converges to the true entropy: $\bar{H}(X) = \frac{1}{n}\log\frac{1}{p(X_1,\ldots,X_n)} \xrightarrow[n\to\infty]{\text{i.p.}} H(X)$.

Example

Consider $X \sim \mathrm{Bern}(p)$. If $X_1, X_2, \ldots, X_n$ are i.i.d., then $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. For example, $p(1,0,1,1,0,1) = p^{\sum x_i} \times (1-p)^{n-\sum x_i} = p^4(1-p)^2$. Clearly, not all sequences are generated equally.

We will see that $p(X_1, X_2, \ldots, X_n)$ is close to $2^{-nH(X)}$ with high probability. That is, the probability $p(X_1, X_2, \ldots, X_n)$ assigned to an observed sequence is close to $2^{-nH(X)}$.

Typical Sequences

Almost all events are almost equally surprising.

$$\Pr\Big\{(X_1, X_2, \ldots, X_n) : p(X_1, X_2, \ldots, X_n) = 2^{-n(H\pm\epsilon)}\Big\} \approx 1,$$

if $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim p(x)$.

We can thus divide the set of all sequences into two classes:

1. **Typical set**, where the probability of each typical sequence is close to $2^{-nH(X)}$.
2. **Atypical set** that contains all the other sequences.
3. Typical set is primarily a theoretical tool that is defined to help prove some theorems, even though its concept is somehow counter-intuitive, as we will see later.

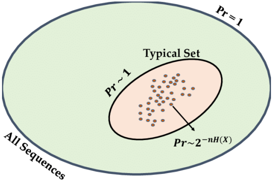

Figure: Typical sequences and typical set.

Asymptotic Equipartition Property (AEP)

Theorem (AEP: Empirical entropy converges to the true entropy.)

If $X_1, X_2, \ldots, X_n$ *are i.i.d. and* $X_i \sim p(x)$, *then*

$$-\frac{1}{n} \log p\left(X_1, X_2, \ldots, X_n\right) \xrightarrow{i.p.} H(X).$$

Proof: by the weak law of large numbers (WLLN), we have

$$-\frac{1}{n} \log p\left(X_1, X_2, \ldots, X_n\right) = -\frac{1}{n} \sum_i \log p\left(X_i\right) \xrightarrow{i.p.} -\mathrm{E}[\log p(X)] = H(X)$$

Example (Sanity check of AEP)

Consider i.i.d. $X_i \sim \text{Bern}(p)$, $i = 1, 2, \ldots n$, let $q = 1 - p$. We have

$p(x_1, \ldots, x_n) = p^{\sum_{i=1}^n x_i} \times q^{n - \sum_{i=1}^n x_i} \xrightarrow{i.p.} p^{np} q^{nq}$.

Since $H(X) = -\log(p^p q^q) \implies -nH(X) = \log(p^{np} q^{nq})$.

This matches the AEP: $p(X^n) \xrightarrow{i.p.} 2^{-nH(X)}$.

Q: AEP is based on the assumption that X^n are i.i.d. How about for non-iid case?

A: Entropy rate of stochastic processes; see the next lecture (Chap 4).

Typical Sequences and Set

Some sequences are "typical" in the sense that their information is about the same as the self-information expected. We define those typical sequences as follows:

Definition (ϵ-typical sequence)

A sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ is an ϵ-typical sequence with respect to $p(x)$ if

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

Definition (ϵ-typical set)

The typical set $A_\epsilon^{(n)}$ is the set containing all ϵ-typical sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

- Intuitively, we would like assign shorter bit strings to the most probable (more 'typical') sequences to reduce the expected length of the code. It turns out that as n grows large, all the probability concentrates on the typical set.
- Typical set is made by all the sequences that are giving us an amount of information close to the average information of the source distribution.
- The most probable sequence usually gives us less information than the average.

Typical Sequences and Set (Cont'd)

Example (Most likely sequence is often not in the typical set, from Wiki)

- For i.i.d. $X_i \sim \text{Bern}(0.9)$, $H(X) = 0.469$. The most likely sequence of outcome is the sequence of all 1's, $(1, 1, ..., 1)$.

$$-\frac{1}{n}\log_2 p\left((x_1, x_2, \ldots, x_n) = (1, 1, \ldots, 1)\right) = -\frac{1}{n}\log_2(0.9^n) = 0.152$$

Hence, for small enough ϵ, all-one sequence is not in the typical set.

- For Bernoulli random variables, the typical set consists of sequences with average numbers of 0s and 1s in n independent trials. Because if a sequence has np 1s and nq 0s for n trails, then $p(x_1, \ldots, x_n) = p^{np}q^{nq} \implies$

$$-\frac{1}{n}\log_2 p\left(x_1, x_2, \ldots, x_n\right) = -p\log p - q\log q = H(X).$$

If $p = 0.9, n = 10$, then the typical set consist of all sequences that have a single 0 in the entire sequence. If $p = 0.5$, then every possible binary sequences belong to the typical set.

Typical Sequences and Set (Cont'd)

Example (The "typicality" is in the sense of *sample entropy close to the true entropy*, rather than "most likely")

A computer program is used to generate a binary sequence of length 10 digits (i.i.d. $X_i \sim \text{Bern}(\frac{1}{3})$). One of the following four sequences is generated from the program. Which one is it?

(a) 0 0 0 0 0 0 0 0 0 0 0 0, $\Pr(\text{a}) = (2/3)^{12} = 7.7 \times 10^{-3}$

(b) 1 0 1 1 0 1 0 1 0 1 0 0, $\Pr(\text{b}) = (2/3)^6 \times (1/3)^6 = 1.2 \times 10^{-4}$

(c) 0 0 0 1 0 0 0 1 0 0 1 0, $\Pr(\text{c}) = (2/3)^9 \times (1/3)^3 = 9.6 \times 10^{-4}$

(d) 1 1 1 1 1 1 1 1 1 1 1 1, $\Pr(\text{d}) = (1/3)^{12} = 1.9 \times 10^{-6}$.

The answer is sequence (c), although sequence (a) has a higher probability of occurrence. An intuition based reasoning is that, since the source outputs are i.i.d., roughly $1/3$ of the 12 digits should be zero and $2/3$ should be one. This is in fact true as the length of the sequence is increasing. Those sequences are called "typical sequences".

Typical Sequences and Set (Cont'd)

- Consider a random source i.i.d. $X_i \sim \text{Bern}(p)$ generating a sequence of length n.
- There are $\binom{n}{np}$ independent sequences that have exactly np ones and the probability of each such sequence is $p^{np}(1-p)^{n(1-p)}$.
- Approximate $\binom{n}{np}$ by using the Stirling's formula: $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$.

$$
\begin{aligned}
\log \binom{n}{np} &= \log \left(\frac{n!}{(np)!(n-np)!} \right) \\
&\approx \log \left(\frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi np}\left(\frac{np}{e}\right)^{np} \sqrt{2\pi n(1-p)}\left(\frac{n(1-p)}{e}\right)^{n(1-p)}} \right) \\
&= -\log\left(\sqrt{2\pi np(1-p)}\right) - n \times [p \log p + (1-p)\log(1-p)]
\end{aligned}
$$

$$
\Longrightarrow \boxed{\binom{n}{np} \approx \frac{2^{nH(p)}}{\sqrt{2\pi np(1-p)}}}
$$

Hence, the number of such sequences increases as $2^{nH(p)}$, but it's a much smaller subset of all possible sequences.

Properties of AEP

Theorem

Let $x^n := (x_1, x_2, \ldots, x_n)$*, we have the following properties of the AEP:*

1. *If* $x^n \in A_\epsilon^{(n)}$ *then* $H(X) - \epsilon \leq -\frac{1}{n}\log p(x^n) \leq H(X) + \epsilon$*.*
2. $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ *for* n *sufficiently large.*
3. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$*.*
4. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for* n *sufficiently large.*

Proof.

1. From the definition of ϵ-typical sequences, if $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then we have

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

Taking the $\log$ of all sides and then dividing by $-n$ gives the 1st property.

Properties of AEP

Proof (Cont.)

2. Since $-\frac{1}{n}\log p(X^n) \xrightarrow{i.p.} H(X)$, for any $\delta \geq 0$, there exists n_0 such that for all $n \geq n_0$,
$$\Pr\left(\left|-\frac{1}{n}\log p(X^n) - H(X)\right| < \epsilon\right) > 1-\delta$$
Finally, we know $\left|-\frac{1}{n}\log p(X^n) - H(X)\right| < \epsilon$ holds iif X^n is ϵ-typical, so we can set $\delta := \epsilon$ to obtain the 2nd property.
3. Since $A_\epsilon^{(n)} \subseteq \mathcal{X}^n$, we have $1 = \sum_{x^n \in \mathcal{X}^n} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)$.
By the definition of ϵ-typical sequences, we have $p(x^n) \geq 2^{-n(H(X)+\epsilon)}$, $\forall x^n \in A_\epsilon^{(n)}$.
Hence, $1 \geq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \geq 2^{-n(H(X)+\epsilon)} \left|A_\epsilon^{(n)}\right|$, which proves the 3rd property.
4. By the 2nd property, if n is sufficiently large, then $\Pr\{A_\epsilon^{(n)}\} > 1-\epsilon$. So, we have
$$1-\epsilon < \Pr\{A_\epsilon^{(n)}\} = \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} \left|A_\epsilon^{(n)}\right|.$$
The last inequality follows from the definition of ϵ-typical sequences.

Properties of AEP (Cont'd)

For small ϵ, we have $|A_\epsilon^{(n)}| \approx 2^{nH(X)}$. Thus, the fraction of sequences that are typical is

$$\rho_n := \frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} \approx \frac{2^{nH(X)}}{|\mathcal{X}|^n} = \frac{2^{nH(X)}}{2^{n\log|\mathcal{X}|}} = 2^{-n(\log|\mathcal{X}|-H(X))}.$$

- For non-uniform distribution: $H(X) < \log|\mathcal{X}|$, $\rho_n \to 0$ as $n \to \infty$.
- For uniform distribution: $H(X) = \log|\mathcal{X}|$, so every sequence is typical.

Everything outside the typical set has a negligible probability.

- $|A_\epsilon^{(n)}|$ **is exponentially small fraction in** n. However, the typical sequences make up most of the probability because $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$.
- In other words, the probability of a generated sequence being in the typical set is high, even though the number of elements in the typical set is much smaller than the total number of possible sequences.
- For n sufficiently large, all the typical sequences have about the same probability $2^{-nH(X)} \to$ "asymptotic equipartition".

High-probability Set

To this end, we know that the $A_\epsilon^{(n)}$ is a fairly small set that has most of the probability.

Q: Is it the smallest set with such a property?

Definition

For $\delta > 0$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set such that $\Pr\left(X^n \in B_\delta^{(n)}\right) \geq 1 - \delta$.

Theorem

Let $\delta < \frac{1}{2}$. For any $\delta' > 0$,

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

for n sufficiently large.

Typical set vs High-probability set.

For sufficiently large n (depending on δ and δ'), $B_\delta^{(n)}$ has at least $2^{n(H-\delta')}$ elements. The ϵ-typical set $A_\epsilon^{(n)}$ has about $2^{n(H\pm\epsilon)}$ elements. Thus, $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ have roughly the same number of elements to first order in the exponent.

Encoding for the Typical Set

The fact that the typical set has probability approaching 1 as n grows large means that we "only need" to care about encoding the sequences in the typical set.
Note that, the number of bits required to encode a set of size $\mathcal{S}$ is $\lceil \log |\mathcal{S}| \rceil$, where $\lceil a \rceil$ outputs the smallest integer number $\geq a$ (ceiling).

Let i.i.d. $X_i \sim p(x)$, $i = 1, \ldots, n$. Consider the following scheme for coding $x^n \in \mathcal{X}^n$.

- First, consider a complete order of all the sequences in the sets $A_\epsilon^{(n)}$ and its complement, according to a certain criterion (e.g., lexicographic order, "ABC, ACB, BAC, BCA, CAB, CBA").
- We use the first bit as an indicator to represent if x^n is typical or not: start with 0 if the sequence is typical, otherwise start with 1.

Encoding for the Typical Set (Cont'd)

- If $x^n \in A_\epsilon^{(n)}$, since $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$, use $n(H(X)+\epsilon)+1$ bits for encoding (correction of 1 bit because of integrality),
- If $x^n \notin A_\epsilon^{(n)}$, use no more than $n \log |\mathcal{X}| + 1$ bits to encode it.

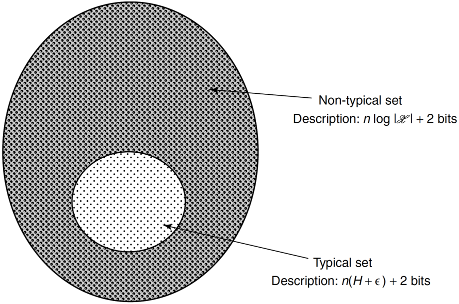

Figure: Encoding for the typical set.

Consequences of AEP

Let $\ell(x^n)$ denote the length of the codeword (a binary string) corresponding to $x^n \in \mathcal{X}^n$. For a sufficiently large n, we have

$$
\begin{aligned}
\mathrm{E}[\ell(x^n)] &\leq P\left(x^n \in A_\epsilon^{(n)}\right) \times (n(H+\epsilon)+2) + P\left(x^n \notin A_\epsilon^{(n)}\right) \times (n \log|\mathcal{X}|+2) \\
&= 2 + P\left(x^n \in A_\epsilon^{(n)}\right) \times (n(H+\epsilon)) + P\left(x^n \notin A_\epsilon^{(n)}\right) \times (n \log|\mathcal{X}|) \\
&\leq 2 + n(H+\epsilon) + \epsilon n \log|\mathcal{X}| =: n(H+\tilde{\epsilon})
\end{aligned}
$$

where $\tilde{\epsilon} = \epsilon(1+\log|\mathcal{X}|) + \frac{2}{n}$ can be arbitrarily small by appropriate choices of ϵ and n.

Theorem (On average $H(X)$ bits are needed to encode X^n per symbol)

Consider i.i.d. $X_i \sim p(x)$, $i = 1, \ldots, n$. *Let* $\epsilon > 0$, *then there exists a code that maps sequences* x^n *into binary strings, such that mapping is one-to-one and* $E\left[\frac{1}{n}\ell(x^n)\right] \leq H(x) + \epsilon$, *for n sufficiently large.*

The above theorem explains the achievability part of the *Source Coding Theorem*: A sequence of symbols can be compressed to a binary string with an average of $H(X)$ bits per symbol. This further reinforces the interpretation of the entropy as the average information content of a random variable (or a random source).

Jointly Typical Sequences

Two sequences x^n and y^n are jointly ϵ-typical if

1. the pair (x^n, y^n) is ϵ-typical with respect to the joint distribution $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$ (i.e., pairwise independence).
2. both x^n and y^n are ϵ-typical w.r.t. their marginal distributions $p(x^n)$ and $p(y^n)$.

The set of all such pairs of sequences (x^n, y^n) is denoted by

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

Joint AEP

Theorem (Joint AEP; see the proof on page 196–198 of the textbook)

Let (X^n, Y^n) *be i.i.d.* $\sim p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. *Then,*

1. $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1$ *as* $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, *then*

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

where the upper bound holds for n sufficiently large.

Implication:

- Typical sets $|X^n| \approx 2^{nH(X)}$ and $|Y^n| \approx 2^{nH(Y)}$.
- Not all pair of typical X^n and typical Y^n are jointly typical: only about $2^{nH(X,Y)}$.
- The probability of any randomly chosen pair is jointly typical is about $\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X;Y)}$

We use the joint AEP and random coding to prove the channel coding theorem (Chap 7).

Thank You!

Email: <zhangy@ucsc.edu>

Homepage: https://people.ucsc.edu/~yzhan419/