Information Theory and Communication
Asymptotic Equipartition Property

Ritwik Banerjee

rbanerjee@cs.stonybrook.edu
Law of Large Numbers

- What happens when we perform the same experiment again and again, ad infinitum?
  - Under the assumption that each run is independent of the other runs.
- The average of the results converges to the expected value.
- There are two versions of this law, weak and strong. Both versions say the following in slightly different ways:

**Law of Large Numbers**

- For an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$, with expected value $E(X_i) = \mu$ for all integer $i > 0$,

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.
$$
The asymptotic equipartition property (AEP) is the information theoretic analog of the law of large numbers in probability theory. It states that for i.i.d. random variables $X_1, X_2, \ldots$,

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{p(X_1, X_2, \ldots, X_n)} = H$$

i.e.,

$$\lim_{n \to \infty} p(X_1, X_2, \ldots, X_n) = 2^{-nH}.$$

What does $H$ denote here?
Asymptotic Equipartition Property

- The asymptotic equipartition property (AEP) is the information theoretic analog of the law of large numbers in probability theory.
- It states that for i.i.d. random variables $X_1, X_2, \ldots$,

$$
\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{p(X_1, X_2, \ldots, X_n)} = H
$$

i.e.,

$$
\lim_{n \to \infty} p(X_1, X_2, \ldots, X_n) = 2^{-nH}.
$$

- What does $H$ denote here?
  - It helps to think of $X_1, X_2, \ldots$ as a time-series of i.i.d. random variables that are coming from a source $X$.
  - $H$ in the above equations denotes the entropy of this source $X$.
- The property can be derived from the weak law of large numbers:

For an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$, with expected value $E(X_i) = \mu$ for all integer $i > 0$,

$$
\lim_{n \to \infty} Pr \left( |\bar{X}_n - \mu| > \epsilon \right) = 0 \forall \epsilon > 0
$$
Convergence of Random Variables

- When $X_1, X_2, \ldots$ is a time-series (from repeatedly performing some experiment), $p(X_1, X_2, \ldots, X_n)$ denotes the probability of observing a sequence.

- Returning to the coin tossing experiment, where $X \in \{0, 1\}$ and the distribution is $Pr(1) = p$, $Pr(0) = q = 1 - p$.

- If $X_1, \ldots, X_n$ are i.i.d., the probability of a sequence $x_1, x_2, \ldots, x_n$ is

$$Pr(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} Pr(X_i = x_i)$$

- E.g., $Pr(101101) = p^4 q^2 = p^4 (1 - p)^2$.

- So clearly, not all of the $2^n$ sequences have the same probability.

- However, we can still predict the probability of observing a particular sequence.

- This is the probability $p(X_1, X_2, \ldots, X_n)$ of the outcomes $X_1, X_2, \ldots, X_n$, where the random variables are all i.i.d. with distribution $p(x)$. 
Convergence of Random Variables

- It is in this *context* that $p(X_1, X_2, \ldots, X_n)$ is close to $2^{-nH}$ with high probability.

- In mathematics, high probability is usually indicated by the use of the word “almost”.

- Thus, almost all event outcome sequences are almost equally surprising, and can be written as

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

- In terms of the coin tosses, what this says is that
  - the number of 1s in a sequence of length $n$ is $np$ with high probability, and
  - the probability of all such sequences is approximately $2^{-nH(p)}$. 

© Ritwik Banerjee
Information Theory and Communication
Convergence of Random Variables

Definition

Given a sequence of random variables \( \{X_i\}_{i=0,1,...} \), there are three types of convergence to a random variable \( X \).

- **Convergence in probability**
  For any \( \epsilon > 0 \), \( Pr \{ |X_n - X| > \epsilon \} \to 0 \) as \( n \to \infty \).

- **Convergence in mean square**
  \[ E(X_n - X)^2 \to 0. \]

- **Convergence with probability 1 (or “almost surely”)**
  \[ Pr \left\{ \lim_{n \to \infty} X_n = X \right\} = 1. \]
Convergence of Random Variables

Figure: Convergence in probability versus almost sure convergence\(^1\).

(a) **Almost sure convergence** says that this curve will eventually, at some finite \(n\), fall entirely within the \(\epsilon\)-band around 0 and from there on, stay within that \(\epsilon\)-band.

(b) **Convergence in probability** says that a large proportion of the sample paths will fall within the \(\epsilon\)-band, and will remain in the band thereafter. We can’t be sure that any particular path will be inside the band, but the fraction tends to 1 as \(n\) keeps increasing.

\(^1\) Source: [https://stats.stackexchange.com/a/11013/12814](https://stats.stackexchange.com/a/11013/12814).
Typical Sets

- Not all sequences of length $n$ have the same probability, but most sequences have a probability close to $2^{-nH}$.
- In other words, we can divide the set of all sequences into two sets,
  - the *typical* set, where the sample entropy is close to the true entropy, and
  - the nontypical set, which contains the other sequences.

Typical Sets

The **typical set** $A_{\epsilon}^{(n)}$ w.r.t. a probability mass function $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ that satisfy

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

- Note that a typical set is defined in terms of the two parameters $n$
  - the sequence length $n$, and
  - $\epsilon$, a positive number that reflects how close the sample entropy is to the true entropy.
Typical Sets

Lemma 1
For any \((x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)},\)

\[
H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon
\]

- Proof is directly from the definition of typical sets, by taking logarithm on the inequalities.
Typical Sets

**Lemma 2**
For any $\epsilon > 0$, there exists an integer $n > 0$ such that

$$Pr \left\{ A^{(n)}_{\epsilon} \right\} > 1 - \epsilon$$

- Proof follows from AEP and using the $\epsilon$-$\delta$ definition of limit.
Typical Sets

Lemma 3

\[ |A^{(n)}_\epsilon| \leq 2^{n(H(X)+\epsilon)} \]

- Generate an inequality based on the fact that the sum of all the probabilities is 1.
- Then use the definition of \( A^{(n)}_\epsilon \).
Typical Sets

Lemma 4
For large enough values of $n$, we have

$$\left| A^{(n)}_\epsilon \right| \geq (1 - \epsilon)2^{n(H(X) - \epsilon)}$$

- Use lemma 2