ECE 7680
Lecture 4 – The Asymptotic Equipartition Property

Objective: The AEP is the weak law of large numbers as applied to the estimation of entropy.

The law of large numbers and AEP

The law of large numbers states that for i.i.d. random variables $X_1, X_2, \ldots, X_n$ the sum

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

is close to its expected value EX. (How close? How can we tell?) The AEP states that for i.i.d. r.v.'s,

$$\frac{1}{n}\log\frac{1}{p(X_1, X_2, \ldots, X_n)}$$

is close to $H(X)$. To put it another way,

$$p(x_1, x_2, \ldots, x_n) \approx 2^{-nH}$$

Example 1 Suppose $X \in \{\mathbf{0}, \mathbf{1}\}$, with $p(\mathbf{1}) = p$, $p(\mathbf{0}) = q$. The probability of the sequence $X_1, X_2, \ldots, X_n$ (iid) is $p^{\sum X_i} q^{n-\sum X_i}$.

Specifically take $p(\mathbf{0}) = 0.7$, $p(\mathbf{1}) = 0.3$, and consider the probabilities of the following sequences ($n = 10$):

Sequence	Probability	number	prob x number
0000000000	.0282	1	0.0282
0000000001	.0121	10	0.121
0000000011	.005	45	0.225
0000000111	.0022	120	0.264
0000001111	.00095	210	0.1995
0000011111	.0004	252	.1008
0000111111	.00017	210	.036
0001111111	.000075	120	.009
0011111111	.000032	45	.00144
0111111111	.0000138	10	.000138
1111111111	.0000059	1	.0000059

Clearly, not all 2^n sequences of length n have the same probability. Those sequences with the number of **1**s approximately np have the highest total probability: they are "typical". This property becomes more pronounced as n increases.

What is the probability $p(X_1, X_2, \ldots, X_n)$ of the outcomes $X_1, X_2, \ldots, X_n$? It turns out that $p(X_1, X_2, \ldots, X_n)$ is close to 2^{-nH}, with high probability. That is,

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

Since the entropy conveys a measure of "surprise" at observing the outcome of a random variable, we can summarize this rule as: "Almost all events are almost equally surprising." That is, most events occur with a probability that is related to the entropy of the ensemble.

In the particular example just considered, we are simply saying that the number of **1**s observed is close to np (with high probability), and all such sequences have roughly the same probability. This is nothing more than the law of large numbers. □

Convergence in probability

Before jumping intro the theorem we need to discuss what it means to converge *in probability*. There are a variety of ways that things can converge in the world of probability. There is convergence *in distribution* (as exhibited by the central limit theorem), there is convergence *almost surely* (which is very strong), there is convergence in *mean square*, and there is convergence *in probability*. We could spend about three weeks working through what each of these mean, and which implies the other. However, for the moment we will simply focus on convergence in probability.

Definition 1 A sequence $X_1, X_2, \ldots, X_n$ is said to **converge in probability** to X if

$$P[|X - X_n| > \epsilon] \to 0 \text{ as } n \to \infty.$$

□

Recalling the definition of convergence of sequences, we can say this as follows: For any $\epsilon > 0$ and for any $\delta > 0$, there is an n_0 such that

$$P[|X - X_n| < \epsilon] > 1 - \delta$$

for all $n > n_0$.

Example 2 Let X_i be a sequence of iid random variables and let

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_n$$

be the sample mean. Let $S = ES$ be the true (ensemble) mean. Then (this again is the WLLN)

$$P[|S_n - S| > \epsilon] \to 0 \text{ as } n \to \infty.$$

More briefly we might write $S_n \to S$ in probability. □

One way of quantifying how we are doing on the convergence is by means of Markov's inequality: For a positive r.v. and any $\delta > 0$,

$$P[X \geq \epsilon] \leq \frac{EX}{\delta}.$$

From this we can derive the **Chebyshev inequality**: for a r.v. Y with mean μ and variance σ^2

$$P[|Y - \mu| > \epsilon] \leq \frac{\sigma^2}{\epsilon^2}.$$

From this we can show convergence of the sample mean (the WLLN).

Now the AEP:

Theorem 1 *(AEP) If $X_1, X_2, \ldots, X_n$ are iid $\sim p(x)$ then*

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

in probability.

Proof Since the X_i are iid, so are the $\log p(X_i)$. By independence and the WLLN,

$$\begin{aligned} -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n} \sum_i \log p(X_i) \\ &\to -E \log p(X) \text{ in probability} \\ &= H(X). \end{aligned}$$

□

The set of sequences that come up most often (according the probability law of the r.v.) are *typical sequences*. The typicality is defined as follows:

Definition 2 The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ with the following property:

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

□

So the typical sequences occur with a probability that is in the neighborhood of 2^{-nH}.

Example 3 Returning to the first example, $H(X) = 0.88129$, and $nH(X) = 8.8129$. Then $2^{-nH(X)} = 0.00223$. Notice that this is very close to the probability with which the sequence with three ones occurs. □

The typical set $A_\epsilon^{(n)}$ has the following properties:

Theorem 2 *1. If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$ *then* $H(X) - \epsilon \leq -\frac{1}{n} p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$.

2. $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ *for* n *sufficiently large.*

3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$*, where* $|A|$ *denotes the number of elements in the set* A.

4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for* n *sufficiently large.*

Interpretations: By (1), nearly all of the elements in the typical set are nearly equiprobable. By (2), the typical set occurs with probability near 1 (that is why it is typical!). By (3) and (4), the number of elements in the typical set is nearly 2^{nH}.

Proof

1. Take $-\frac{1}{n}\log_2$ of the definition of the typical set:

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

2. By the definition of typical sets, the AEP, and the definition of convergence in probability,

$$\Pr\{(X_1, \ldots, X_n) | (X_1, \ldots, X_n) \in A_\epsilon^{(n)}\} \to 1 \text{ as } n \to \infty$$

That is, for any $\delta > 0$ there is an n_0 s.t. for all $n \geq n_0$ we have

$$\Pr\{|-\frac{1}{n}\log p(X_1, \ldots, X_n) - H(X)| < \epsilon\} > 1 - \delta.$$

Set $\delta = \epsilon$, and we obtain part (2) of the theorem.

3. To prove (3),

$$\begin{aligned} 1 &= \sum_{x \in \mathcal{X}^n} p(\mathbf{x}) \\ &\geq \sum_{x \in A_\epsilon^{(n)}} p(\mathbf{x}) \\ &\geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(x)+\epsilon)} \\ &= 2^{-n(H(x)+\epsilon)} |A_\epsilon^{(n)}|. \end{aligned}$$

4. For part (4), for sufficiently large n, $\Pr\{A_\epsilon^{(n)}\} > 1-\epsilon$, so

$$\begin{aligned} 1-\epsilon &< \Pr\{A_\epsilon^{(n)}\} \\ &\le \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= 2^{-n(H(X)-\epsilon)}|A_\epsilon^{(n)}|. \end{aligned}$$

□

Where does this lead: data compression

Typical sequences occur most of the time. Therefore, one way to perform data compression is to only provide efficient codes for the typical sequences, and other (less efficient) codes for the non-typical sequences. Since there are $\le 2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, then we can encode every one of them with no more than $n(H+\epsilon)+1$ bits. Let us prefix each typical-sequence code with a 0, so the total length of codes for typical sequences is $\le n(H+\epsilon)+2$.

Each sequence not in $A_\epsilon^{(n)}$ can be indexed with not more than $n\log|\mathcal{X}|+1$ bits. If these codes are prefixed with a 1, then we have a code for all sequences in $\mathcal{X}^n$. Even though we are doing a brute-force enumeration of the atypical set (overlooking the fact that we don't need to do those codes that are already in the typical set), we can still do a good job (on the average).

Notation: let x^n denote the sequence $x_1, x_2, \ldots, x_n$.

Theorem 3 *Let X^n be i.i.d., and let $\epsilon > 0$. Then there exists a code which maps sequences x^n into binary strings such that the mapping is one-to-one (lossless) and the average codeword length satisfies*

$$E[\frac{1}{n}l(X^n)] \le H(X) + \epsilon$$

for n sufficiently large.

This theorem does *not* go so far yet as to say that this is about the best that we can do.

Proof The expected length of the codeword is

$$\begin{aligned} El(X^n) &= \sum_{x^n} p(x^n)l(x^n) = \sum_{x^n\in A_\epsilon^{(n)}} p(x^n)l(x^n) + \sum_{x^n\in A_\epsilon^{(n)c}} l(x^n)p(x^n) \\ &\le \sum_{x^n\in A_\epsilon^{(n)}} p(x^n)[n(H+\epsilon)+2] + \sum_{x^n\in A_\epsilon^{(n)c}} (n\log|\mathcal{X}|+2)p(x^n) \\ &= \Pr(A_\epsilon^{(n)})[n(H+\epsilon)+2] + \Pr(A_\epsilon^{(n)c})(n\log|\mathcal{X}|+2) \\ &\le n(H+\epsilon) + \epsilon n(\log|\mathcal{X}|) + 2 \\ &= n(H+\epsilon') \end{aligned}$$

where

$$\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$$

can be made as small as desired by choice of ϵ followed by choice of n.

□

High probability sets and the typical set

We introduce here some notation that is useful in later chapters.

Example 4 We begin with an example. Suppose a_n is a sequence defined as $a_n = e^{3n+1}$ and b_n is a sequence $b_n = e^{3n}$. Then

$$\log \frac{a_n}{b_n} = \log e,$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

Thus, even though a_n and b_n are different at every point (by a factor of e, to be precise), the log of the ratio doesn't grow too fast. □

Example 5 We present another example of the same concept. Suppose $a_n = e^{3n+\sqrt{n}}$ and $b_n = e^{3n}$. Then

$$\log \frac{a_n}{b_n} = \sqrt{n}$$

and

$$\lim_{n\to\infty} \log \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0.$$

Here, even though the factor between a_n and b_n is increasing, it is increasing slowly. □

Example 6 Now let $a_n = e^{4n}$ and $b_n = e^{3n}$. Then

$$\log \frac{a_n}{b_n} = n,$$

and

$$\lim_{n\to\infty} \log \frac{a_n}{b_n} = 1.$$

So the ratio increases too fast. □

By these examples, we are in a better position to understand the following definition.

Definition 3 The notation $a_n \doteq b_n$ means

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

That is, a_n and b_n **agree to first order in the exponent**. □

Now, the typical set $A_\epsilon^{(n)}$ is a fairly small set that contains most of the probability, but it is not clear if it is the smallest such set. We will show that any small but highly probable set must contain a significant overlap with the typical set.

Definition 4 Let $B_\delta^{(n)} \in \mathcal{X}^n$ be any set with

$$\Pr(B_\delta^{(n)}) \geq 1 - \delta.$$

$B_\delta^{(n)}$ may be viewed as a "high probability set." □

Then we have the following theorem:

Theorem 4 *Let X_i be i.i.d. and $\sim p(x)$. For $\delta < 1/2$ and any $\delta' > 0$, if $(Pr)\{B_\delta^{(n)}\} \geq 1-\delta$ then*

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

for n sufficiently large. Thus, $B_\delta^{(n)}$ must have at least 2^{nH} elements in it, to first order in the exponent. But $A_\epsilon^{(n)}$ has about $2^{nH\pm\epsilon}$ elements in it, so $A_\epsilon^{(n)}$ is about the same size as the smallest high probability set:

$$|B_\delta^{(n)}| \doteq |A_\epsilon^{(n)}| \doteq 2^{nH}.$$

To contrast, consider X_i be Bernoulli(.9). Then a typical set contains sequences in which the proportion of 1's is close to .9. But it does *not* contain the single most probable sequence, the sequence of all 1s. The set $B_\delta^{(n)}$ (the high probability set) will include all the most probable sequences, including the sequence of all 1s. By the theorem, A and B must both contain the sequences that are about 90% 1s, and that the two sets are almost equal in size.