Introduction to AEP

In information theory, the asymptotic equipartition property (AEP) is the analog of the law of large numbers. This law states that for independent and identically distributed (i.i.d.) random variables:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n->\infty]{} EX$$

Similarly, the AEP states that:

$$\frac{1}{n}\log\frac{1}{p(X_1, X_2,...X_n)} \xrightarrow[n->\infty]{} H$$

Where $p(X_1,X_2,\ldots X_n)$ is the probability of observing the sequence $X_1,X_2,\ldots X_n$. Thus, the probability assigned to an observed sequence will be close to 2^{-nH} (from the definition of entropy).

Consequences

We can divide the set of all sequences in to two sets, the typical set, where the sample entropy is close to the true entropy, and the non-typical set, which contains the other sequences. The importance of this subdivision is that <u>any property that is proven for the typical sequences will then be true with high probability and will determine the average behavior of a large sample</u> (i.e. a sequence of a large number of random variables).

For example, if we consider a random variable $X \in \{0,1\}$ having a probability mass function defined by $p(1)=p$ and $p(0)=q$, the probability of a sequence $\{x_1,x_2,\ldots x_n\}$ is:

$$\prod_{i=1}^{n} p(x_i)$$

For example, the probability of the sequence (1,0,1,1,0,1) is p^4q^2. Clearly, it is not true that all 2^n sequences of length n have the same probability. In this example, we can say that the number of 1's in the sequence is close to np.

Convergence of Random Variables

Definition: Given a sequence of random variables $X_1, X_2, \ldots X_n$, we say that the sequence $X_1, X_2, \ldots X_n$, converges to a random variable X:

1. *In probability* if for every $\varepsilon>0$, $\Pr\{|X_n-X|>\varepsilon\} \to 0$
2. *In mean square* if $E(X_n-X)^2 \to 0$
3. *With probability 1* (also called almost surely) if $\Pr\{\lim_{n\to\infty} X_n=X\}=1$

The AEP

Theorem (AEP): If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

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

in probability.

Proof: Functions of independent random variables are also independent random variables. Thus, since the X_i are i.i.d., so are $\log p(X_i)$. Hence by the law of large numbers:

$$\begin{aligned}\frac{1}{n}\log\frac{1}{p(X_1, X_2, \ldots X_n)} &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i)\\ &\longrightarrow -E\log p(X) \qquad \text{in probability}\\ &= H(X)\end{aligned}$$

Typical Set

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

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

As a consequence of the AEP, we can show that the set $A_\varepsilon^{(n)}$ has the following properties:

Theorem:

1. If $(x_1,x_2,\ldots x_n) \in A_\varepsilon^{(n)}$, then $H(X)-\varepsilon \le -1/n \log p(x_1,x_2,\ldots x_n) \le H(X)+\varepsilon$
2. $\Pr\{A_\varepsilon^{(n)}\} > 1-\varepsilon$ for n sufficiently large
3. $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$, where $|A|$ denotes the number of elements in the set A
4. $|A_\varepsilon^{(n)}| \ge (1-\varepsilon)2^{n(H(X)-\varepsilon)}$ for n sufficiently large

Thus, the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly 2^{nH}

Typical Set

Proof: The proof of property 1 is immediate from the definition of $A_\varepsilon^{(n)}$. The second property follows directly from Theorem AEP. In fact, from (1) and the definition of convergence in probability, we can say that for any $\delta > 0$, there exists an n_0 such that for all $n \ge n_0$, we have:

$$\Pr\left\{\left|-\frac{1}{n}\log p(X_1, X_2, \ldots X_n) - H(X)\right| < \varepsilon\right\} > 1-\delta$$

Setting $\delta = \varepsilon$, we obtain the second part of the theorem, since the sequence that satisfy (1) is by definition a sequence belonging to $A_\varepsilon^{(n)}$. Hence, the probability of the event $(X_1, X_2, \ldots, X_n) \in A_\varepsilon^{(n)}$ tends to 1 as $n \to \infty$. The identification of $\delta = \varepsilon$ will conveniently simplify notation later. To prove property (3), we write:

$$1 = \sum_{x \in \chi^n} p(x) \ge \sum_{x \in A_\varepsilon^{(n)}} p(x)$$

$$\ge \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)} |A_\varepsilon^{(n)}|$$

Where the second inequality follows from the definition of typical set.

Typical Set

Finally, for n sufficiently large, the second property states that $\Pr\{A_\varepsilon^{(n)}\}>1-\varepsilon$, so that:

$$1-\varepsilon < \Pr\left\{A_\varepsilon^{(n)}\right\} \leq \sum_{x\in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)} \mid A_\varepsilon^{(n)} \mid$$

Where the second inequality follows from the definition of typical set.

Data Compression

Let $X_1,X_2,\ldots X_n$ be i.i.d. random variables $\sim p(x)$. We wish to find short descriptions for such sequences of random variables. We divide all sequences in χ^n into two sets: $A_\varepsilon^{(n)}$ and its complement $A_\varepsilon^{(n)c}$.

We order all elements in each set according to some order, then we can represent each sequence of $A_\varepsilon^{(n)}$ by giving the index of the sequence in the set. Since there are $\leq 2^{n(H+\varepsilon)}$ sequences in $A_\varepsilon^{(n)}$ because of property 3, we can use no more than $n(H+\varepsilon)+1$ bits (the extra bit because it may not be an integer). We prefix all these sequences by 0, giving a total length of $n(H+\varepsilon)+2$ bits.

Similarly, we can index each sequence in $A_\varepsilon^{(n)c}$ by using not more than $n \log|\chi|+1$ bits. Prefixing these indices by 1, we have a code for all the sequences in χ^n.

Data Compression

The coding scheme has the following features:

1. The code is one-to-one decodable, using the initial bit to indicate the length of the codeword following
2. We have used a brute force enumeration of $A_\varepsilon^{(n)c}$ without taking into account that the number of elements is less than the number of elements in χ^n
3. The typical sequences have short description of length $\approx$ nH.

We will use notation x^n to denote the sequence $x_1, x_2, \ldots x_n$. Let $l(x^n)$ be the length of the codeword corresponding to x^n. If n is sufficiently large so that $\Pr\{A_\varepsilon^{(n)}\} \geq (1-\varepsilon)$ (property 2), then the expected length of the codeword is:

Data Compression

$$E(l(X^n)) = \sum_{x^n} p(x^n) l(x^n)$$

$$= \sum_{x \in A_\varepsilon^{(n)}} p(x^n) l(x^n) + \sum_{x \in A_\varepsilon^{(n)c}} p(x^n) l(x^n)$$

$$\leq \sum_{x \in A_\varepsilon^{(n)}} p(x^n)[n(H+\varepsilon)+2] + \sum_{x \in A_\varepsilon^{(n)c}} p(x^n)(n \log|\chi| + 2)$$

$$= \Pr\{A_\varepsilon^{(n)}\}[n(H+\varepsilon)+2] + \Pr\{A_\varepsilon^{(n)c}\}(n \log|\chi| + 2)$$

$$\leq n(H+\varepsilon) + \varepsilon n(\log|\chi|) + 2$$

$$= n(H+\varepsilon')$$

Where $\varepsilon' = \varepsilon + \varepsilon \log|\chi| + 2/n$ can be made arbitrarly small. Hence we have proven the following theorem:

Average Code-length

Theorem: Let X^n be i.i.d. ~p(x). Let $\varepsilon>0$. Then there exists a code which maps sequences x^n of length n into arbitrary strings such that the mapping is one-to-one (and therefore invertible) and:

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

For n sufficiently large.

Thus we can represent sequences X^n using nH(X) bits on the average.

High Probability Set and Typical Set

From the definition of $A_\varepsilon^{(n)}$ it is clear that this is a fairly small set that contains most of the probability . But from the definition it is not clear whether it is the smallest such set. We will prove that the typical set has essentially the same number of elements as the smallest set, to first order in the exponent.

Definition: For each n=1,2,…, let $B_\delta^{(n)} \subset \chi^n$ be any set with $\Pr\{B_\delta^{(n)}\} \ge 1-\delta$. We argue that $B_\delta^{(n)}$ must have significant intersection with $A_\varepsilon^{(n)}$, and therefore must have about as many elements.

Theorem: Let X^n be i.i.d. ~p(x). For $\delta<1/2$ and any $\delta'>0$, if $\Pr\{B_\delta^{(n)}\} \ge 1-\delta$, then $1/n \log|B_\delta^{(n)}| > H-\delta'$ for n sufficiently large.

Thus, $B_\delta^{(n)}$ must have at least 2^{nH} elements. But $A_\varepsilon^{(n)}$ has $2^{n(H\pm\varepsilon)}$ elements. Therefore $A_\varepsilon^{(n)}$ is about the same size of the smallest high probability set.

Equality to the First Order

Definition: The notation $a \doteq b$ means:

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

This means that they are equal to the first order exponent.

The above results lead to the following conclusion: If $\delta_n \to 0$ and $\varepsilon_n \to 0$, then: $|B_{\delta n}^{(n)}| \doteq |A_{\varepsilon n}^{(n)}| \doteq 2^{nH}$

Example

To illustrate the difference between the two sets, let us consider a Bernoulli sequence $X_1, X_2, \ldots X_n$ with parameter 0.9. A Bernoulli(θ) random variable is a binary random variable that takes on the value 1 with probability θ. Typical sequences in this case are sequences in which the proportion of 1's is close to θ.

However, this does not include the most likely single sequence, which is the sequence of all 1's. The set $B_{\delta n}^{(n)}$ includes all the most probable sequences, and hence it includes also this sequence. The theorem implies that both A and B contains the sequences that have about 90% of 1's and the two sets are almost equal in size.

Note that the sequences belonging to $A_{\varepsilon}^{(n)}$ have probability near to one but they are not a single sequence. In fact, if we take a single sequence composed by 90% of 1's, this single one it is not the most probable, because there are many of these sequences with the 1's in different positions.