Asymptotic Equipartition Property

This laboratory is intended to shed insight on the distribution of sequences of randomly distributed random variables.

Here is a precise statement of the Asymptotic Equipartition Property (Cover and Thomas, 1991):

If \( X_1, X_2, \ldots \) are independent, identically distributed random variables, and each \( X_j \) has probability mass function \( p(x) \), then

\[
-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \quad \text{in probability},
\]

where \( H(X) \) is the entropy of \( X \), defined as

\[
H(X) = -\sum_x p(x) \log p(x).
\]

This is really just a result of the weak law of large numbers, since

\[
-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_j \log p(X_j) \to E(-\log p(X)) \quad \text{in probability},
\]

and \( E(-\log p(X)) \) is exactly the entropy \( H(X) \).

Just a reminder: a sequence of random variables \( Y_1, \ldots, Y_n \) is said to convergence in probability to \( Y \) if, for each \( \epsilon > 0 \), \( \lim \mathbb{P}(|Y_n - Y| >= \epsilon) = 0 \). In our case, \( Y_n = -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \), and \( Y \) is a constant, namely, \( Y = H(X) \).

Now we try to get an intuitive idea of the AEP.

(1) Pick a discrete type of random variable, e.g., Bernoulli, Binomial, Poisson, Geometric, Hypergeometric distribution, etc.

Let \( p(x) = P(X = x) \) be the probability mass function of this type of random variable. Compute \( H(X) \) for a random variable \( X \) distributed according to the probability mass function that you choose. Once you have picked the distribution, then \( H(X) \) is no longer random, but is just a constant. Generate 1000 samples of your random variable, written as \( x_1, \ldots, x_{1000} \). For each \( n \) from 1 to 1000, let \( a_n = -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \). Does \( a_n \) seem to approach \( H(X) \)? Why or why not?

(2) Now let us investigate the idea of convergence in probability by running the experiment from question (1) several times. Fix a small value of \( \epsilon \), say \( \epsilon = 1/10 \). Pick a big \( n \) value, e.g., \( n = 1000 \). Setup an array called \text{success}[1,2,\ldots,n].

Initialize your array of successes to all zeroes. Run some trials. E.g., run \( m = 200 \) trials. Each trial consists of the following steps:

1. Generate a random sequence \( x_1, \ldots, x_n \) according to the distribution you selected.

2. For each \( j \) with \( 1 \leq j \leq n \), let \( y_j = -\frac{1}{j} \log p(x_1, \ldots, x_j) \).

3. For each \( j \) with \( 1 \leq j \leq n \), if \( |y_j - H(X)| < \epsilon \), then this is a “success”, so increment the value of \text{success}[j].

Now define \( b_j = \text{success}[j]/m \), i.e., \( b_j \) is a good approximation to \( P(| - (1/j) \log p(X_1, \ldots, X_j) - H(X)| < \epsilon) \). We should see that the probabilities \( b_1, b_2, \ldots, b_{1000} \) are approaching 1, so that the probability of success is tending toward 1, and the probability of failing is tending toward 0, by the AEP.
To what extent do the results of your simulation agree with the AEP? What if you change the value of $\epsilon$? How do the values of $n$ and $m$ need to change, as $\epsilon$ changes?

Consider a probability mass function $p(x)$. Define the “typical set” (the notation is again from Cover and Thomas) $A_\epsilon^{(n)}$ as the set of sequences $(x_1, \ldots, x_n)$ such that

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

(The “typical set” depends on the choice of probability distribution, i.e., the typical set is different for Binomial distributions as compared to Poisson distributions or Geometric distributions.) Thus, the probability of each sequence $(x_1, \ldots, x_n)$ in the “typical set” $A_\epsilon^{(n)}$ is approximately $2^{-nH(X)}$, i.e., each sequence in the typical set is almost equally probable. It is a simple mathematical exercise (see p. 52 of Cover and Thomas) to see that there are approximately $2^{nH(X)}$ sequences in the typical set. Based on these two facts, we might guess that the probability of the typical set is nearly 1, i.e., that a randomly selected sequence is almost always located in the typical set.

Indeed, for $\epsilon$ fixed, the probability of a random chosen sequence being located in the typical set will exceed $1 - \epsilon$ for sufficiently large $n$.

(3) Fix a small value of $\epsilon$, say $\epsilon = 1/10$. Pick a big $n$ value, e.g., $n = 1000$. Setup an array called `success[1,2,\ldots,n]`. Initialize your array of successes to all zeroes. Run some trials. E.g., run $m = 200$ trials. Each trial consists of the following steps:

1. Generate a random sequence $x_1, \ldots, x_n$ according to the distribution you selected.

2. For each $j$ with $1 \leq j \leq n$, see if $x_1, \ldots, x_j$ is in the typical set $A_\epsilon^{(j)}$. If so, then this is a “success”, so increment the value of `success[j]`.

Now define $b_j = \text{success}[j]/m$, i.e., $b_j$ is a good approximation to $P(A_\epsilon^{(j)})$. We should see that the probabilities $b_1, b_2, \ldots, b_{1000}$ will exceed $1 - \epsilon$ for $n$ sufficiently large.

Again, what if you change $\epsilon$—how do the values $n$ and $m$ need to change, for $n$ to be sufficiently large so that $P(A_\epsilon^{(j)})$ exceeds $1 - \epsilon$?

Note: A classic observation about data compression is the following: Since there are only a few strings in the “typical set”, these strings can be just be written without compression, which requires $n$ characters. There are lots and lots of strings that are not in the typical set, but the complement of the typical set has very low probability.

(4) (motivated by Cover and Thomas, p. 57) Pick a discrete type of random variable. Let $p(x) = P(X = x)$ be the probability mass function of this type of random variable. Can you compute $\lim_{n \to \infty} [p(X_1, X_2, \ldots, X_n)]^{1/n}$ for the distribution that you chose? Can you generalize?

Run some simulations and see if your hypothesis holds.

(5) What can you say about the distribution of $(X_1 X_2 X_3 \cdots X_n)^{1/n}$ for large $n$?

Again, run some simulations to test your hypothesis.

(6) Are the two results in questions (4) and (5) related? Why?
(7) (Inspired by M. J. Willems) Whenever Alice conducts a random experiment, she flips a fair coin $n$ times, for some large $n$, and records the results (a sequence of $n$ Heads and Tails altogether). Whenever Bob conducts a random experiment, he rolls a fair dice $n$ times, for some large $n$. For each 1 or 2, he records “Heads”; for each 3, 4, 5, or 6, he records “Tails”. Suppose that you are handed a record of an experiment but do not know who conducted it. How should you guess, so that your probability of choosing the correct person is maximized? Now test your strategy, namely:

Generate 1000 random experiments, using Alice’s methodology of coin flipping. What percentage of the time would you have guessed (correctly) that the experiment was conducted by Alice?

Next, generate 1000 random experiments, using Bob’s style of dice rolling. What percentage of the time would you have guessed (correctly) that the experiment was conducted by Bob?