Elements of Information Theory
Thomas M. Cover, Joy A. Thomas
Copyright © 1991 John Wiley & Sons, Inc.
Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 3

The Asymptotic Equipartition Property

In information theory, the analog of the law of large numbers is the Asymptotic Equipartition Property (AEP). It is a direct consequence of the weak law of large numbers. The law of large numbers states that for independent, identically distributed (i.i.d.) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to its expected value EX for large values of n. The AEP states that $\frac{1}{n}\log\frac{1}{p(X_1, X_2, \ldots, X_n)}$ is close to the entropy H, where $X_1, X_2, \ldots, X_n$ are i.i.d. random variables and $p(X_1, X_2, \ldots, X_n)$ is the probability of observing the sequence $X_1, X_2, \ldots, X_n$. Thus the probability $p(X_1, X_2, \ldots, X_n)$ assigned to an observed sequence will be close to 2^{-nH}.

This enables us to divide the set of all sequences into 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. Most of our attention will be on the typical sequences. Any property that is proved for the typical sequences will then be true with high probability and will determine the average behavior of a large sample.

First, an example. Let the random variable $X \in \{0, 1\}$ have a probability mass function defined by $p(1) = p$ and $p(0) = q$. If $X_1, X_2, \ldots, X_n$ are i.i.d. according to $p(x)$, then 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 $p^{\sum X_i} q^{n - \sum X_i} = p^4 q^2$. Clearly, it is not true that all 2^n sequences of length n have the same probability.

However, we might be able to predict the probability of the sequence that we actually observe. We ask for the probability $p(X_1, X_2, \ldots, X_n)$ of the outcomes $X_1, X_2, \ldots, X_n$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. This is insidiously self-referential, but well defined nonetheless. Apparently, we are asking for the probability of an event drawn according to the same probability distribution. Here it turns out that $p(X_1, X_2, \ldots, X_n)$ is close to 2^{-nH} with high probability.

We summarize this by saying, "Almost all events are almost equally surprising." This is a way of saying that

$$\Pr\{(X_1, X_2, \ldots, X_n): p(X_1, X_2, \ldots, X_n) = 2^{-n(H \pm \epsilon)}\} \approx 1, \qquad (3.1)$$

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

In the example just given, where $p(X_1, X_2, \ldots, X_n) = p^{\sum X_i} q^{n - \sum X_i}$, we are simply saying that the number of 1's in the sequence is close to np (with high probability), and all such sequences have (roughly) the same probability $2^{-nH(p)}$.

3.1 THE AEP

The asymptotic equipartition property is formalized in the following theorem:

Theorem 3.1.1 (*AEP*): *If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then*

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \qquad \textit{in probability.} \qquad (3.2)$$

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 weak law of large numbers,

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \sum_i \log p(X_i) \qquad (3.3)$$

$$\to -E \log p(X) \quad \text{in probability} \qquad (3.4)$$

$$= H(X), \qquad (3.5)$$

which proves the theorem. □

Definition: 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:

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}. \qquad (3.6)$$

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

Theorem 3.1.2:

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

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.*

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}.

Proof: The proof of property (1) is immediate from the definition of $A_\epsilon^{(n)}$. The second property follows directly from Theorem 3.1.1, since the probability of the event $(X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. Thus for any $\delta > 0$, there exists an n_0, such that for all $n \geq n_0$, we have

$$\Pr\left\{\left| -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| < \epsilon\right\} > 1 - \delta . \tag{3.7}$$

Setting $\delta = \epsilon$, we obtain the second part of the theorem. Note that we are using ϵ for two purposes rather than using both ϵ and δ. The identification of $\delta = \epsilon$ will conveniently simplify notation later.

To prove property (3), we write

$$1 = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \tag{3.8}$$

$$\geq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} p(\mathbf{x}) \tag{3.9}$$

$$\geq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \tag{3.10}$$

$$= 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}| , \tag{3.11}$$

where the second inequality follows from (3.6). Hence

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)} . \tag{3.12}$$

Finally, for sufficiently large n, $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$, so that

$$1 - \epsilon < \Pr\{A_\epsilon^{(n)}\} \tag{3.13}$$

$$\leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \tag{3.14}$$

$$= 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}| , \tag{3.15}$$

where the second inequality follows from (3.6). Hence

$$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}. \tag{3.16}$$

which completes the proof of the properties of $A_\epsilon^{(n)}$. □

3.2 CONSEQUENCES OF THE AEP: DATA COMPRESSION

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables drawn from the probability mass function $p(x)$. We wish to find short descriptions for such sequences of random variables. We divide all sequences in $\mathcal{X}^n$ into two sets: the typical set $A_\epsilon^{(n)}$ and its complement, as shown in Figure 3.1.

We order all elements in each set according to some order (say lexicographic order). Then we can represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set. Since there are $\leq 2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, the indexing requires no more than $n(H+\epsilon)+1$ bits. (The extra bit may be necessary because $n(H+\epsilon)$ may not be an integer.) We prefix all these sequences by a 0, giving a total length of $\leq n(H+\epsilon)+2$ bits to represent each sequence in $A_\epsilon^{(n)}$. See Figure 3.2.

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

Note the following features of the above coding scheme.

- The code is one-to-one and easily decodable. The initial bit acts as a flag bit to indicate the length of the codeword that follows.
- We have used a brute force enumeration of the atypical set $A_\epsilon^{(n)c}$ without taking into account the fact that the number of elements in $A_\epsilon^{(n)c}$ is less than the number of elements in $\mathcal{X}^n$. Surprisingly, this is good enough to yield an efficient description.
- The typical sequences have short descriptions of length $\approx nH$.

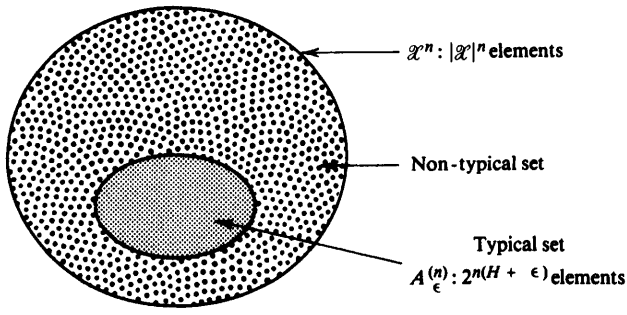


Figure 3.1. Typical sets and source coding.

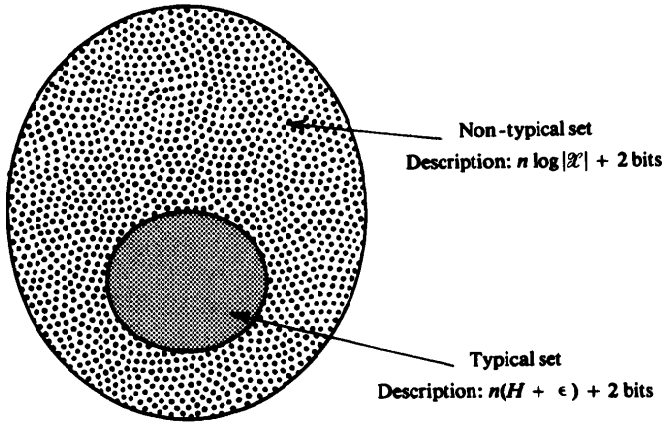


Figure 3.2. Source code using the typical set.

We will use the notation x^n to denote a 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 $\Pr\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$, then the expected length of the codeword is

$$E(l(X^n)) = \sum_{x^n} p(x^n) l(x^n) \tag{3.17}$$

$$= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n) l(x^n) \tag{3.18}$$

$$\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)[n(H + \epsilon) + 2] + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n)(n \log |\mathcal{X}| + 2) \tag{3.19}$$

$$= \Pr\{A_\epsilon^{(n)}\}[n(H + \epsilon) + 2] + \Pr\{A_\epsilon^{(n)c}\}(n \log |\mathcal{X}| + 2) \tag{3.20}$$

$$\leq n(H + \epsilon) + \epsilon n(\log |\mathcal{X}|) + 2 \tag{3.21}$$

$$= n(H + \epsilon'), \tag{3.22}$$

where $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small by an appropriate choice of ϵ followed by an appropriate choice of n. Hence we have proved the following theorem.

Theorem 3.2.1: *Let X^n be i.i.d. $\sim p(x)$. Let $\epsilon > 0$. Then there exists a code which maps sequences x^n of length n into binary strings such that the mapping is one-to-one (and therefore invertible) and*

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \epsilon , \tag{3.23}$$

for n sufficiently large.

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

3.3 HIGH PROBABILITY SETS AND THE TYPICAL SET

From the definition of $A_\epsilon^{(n)}$, it is clear that $A_\epsilon^{(n)}$ 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 $n = 1, 2, \ldots$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be any set with

$$\Pr\{B_\delta^{(n)}\} \geq 1 - \delta . \tag{3.24}$$

We argue that $B_\delta^{(n)}$ must have significant intersection with $A_\epsilon^{(n)}$ and therefore must have about as many elements. In problem 7, we outline the proof of the following theorem:

Theorem 3.3.1: *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\Pr\{B_\delta^{(n)}\} > 1 - \delta$, then*

$$\frac{1}{n} \log|B_\delta^{(n)}| > H - \delta' \quad \text{for } n \text{ sufficiently large} . \tag{3.25}$$

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

We will now define some new notation to express equality to first order in the exponent.

Definition: The notation $a_n \doteq b_n$ means

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0 . \tag{3.26}$$

Thus $a_n \doteq b_n$ implies that a_n and b_n are equal to the first order in the exponent.

We can now restate the above results as

$$|B_\delta^{(n)}| \doteq |A_\epsilon^{(n)}| \doteq 2^{nH} . \tag{3.27}$$

To illustrate the difference between $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$, let us consider a Bernoulli sequence $X_1, X_2, \ldots, X_n$ with parameter $p = 0.9$. (A Bernoulli(θ) random variable is a binary random variable with takes on the value 1 with probability θ.) The typical sequences in this case are the sequences in which the proportion of 1's is close to 0.9. However, this does not include the most likely single sequence, which is the sequence of all 1's. The set $B_\delta^{(n)}$ includes all the most probable sequences, and therefore includes the sequence of all 1's. Theorem 3.3.1 implies that $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ must both contain the sequences that have about 90% 1's and the two sets are almost equal in size.

SUMMARY OF CHAPTER 3

AEP (*"Almost all events are almost equally surprising"*): Specifically, if $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability} . \tag{3.28}$$

Definition: The *typical set* $A_\epsilon^{(n)}$ is the set of sequences $x_1, x_2, \ldots, x_n$ satisfying:

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)} . \tag{3.29}$$

Properties of the typical set:

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

Definition: $a_n \doteq b_n$ means $\frac{1}{n} \log \frac{a_n}{b_n} \to 0$ as $n \to \infty$.

Definition: Let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set such that $\Pr\{B_\delta^{(n)}\} \ge 1 - \delta$, where $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim p(x)$.

Smallest probable set: For $\delta < \frac{1}{2}$,

$$|B_\delta^{(n)}| \doteq 2^{nH} . \tag{3.30}$$

PROBLEMS FOR CHAPTER 3

1. *Markov's inequality and Chebyshev's inequality.*
   (a) (*Markov's inequality*) For any non-negative random variable X and any $\delta > 0$, show that

$$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta} . \tag{3.31}$$

   Exhibit a random variable that achieves this inequality with equality.
   (b) (*Chebyshev's inequality*) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y - \mu)^2$, show that for any $\epsilon > 0$,

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} . \tag{3.32}$$

   (c) (*The weak law of large numbers*) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$ be the sample mean. Show that

$$\Pr\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} . \tag{3.33}$$

   Thus $\Pr\{|\bar{Z}_n - \mu| > \epsilon\} \to 0$ as $n \to \infty$. This is known as the weak law of large numbers.

2. *An AEP-like limit.* Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find

$$\lim_{n \to \infty} [p(X_1, X_2, \ldots, X_n)]^{1/n} .$$

3. *The AEP and source coding.* A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.
   (a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.
   (b) Calculate the probability of observing a source sequence for which no codeword has been assigned.
   (c) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (b).

4. *Products.* Let

$$X = \begin{cases} 1, & \frac{1}{2} \\ 2, & \frac{1}{4} \\ 3, & \frac{1}{4} \end{cases}$$

Let $X_1, X_2, \ldots$ be drawn i.i.d. according to this distribution. Find the limiting behavior of the product

$$(X_1 X_2 \cdots X_n)^{1/n}.$$

5. *AEP.* Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x)$, $x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.
   (a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.
   (b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n} \log \frac{q(X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favoring q are exponentially small when p is true.

6. *Random box size.* An n-dimensional rectangular box with sides $X_1, X_2, X_3, \ldots, X_n$ is to be constructed. The volume is $V_n = \prod_{i=1}^{n} X_i$. The edge length l of a n-cube with the same volume as the random box is $l = V_n^{1/n}$. Let $X_1, X_2, \ldots$ be i.i.d. uniform random variables over the unit interval $[0, 1]$. Find $\lim_{n \to \infty} V_n^{1/n}$, and compare to $(EV_n)^{1/n}$. Clearly the expected edge length does not capture the idea of the volume of the box.

7. *Proof of Theorem 3.3.1.* Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Let $B_\delta^{(n)} \subset \mathcal{X}^n$ such that $\Pr(B_\delta^{(n)}) > 1 - \delta$. Fix $\epsilon < \frac{1}{2}$.
   (a) Given any two sets A, B such that $\Pr(A) > 1 - \epsilon_1$ and $\Pr(B) > 1 - \epsilon_2$, show that $\Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2$. Hence $\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \epsilon - \delta$.
   (b) Justify the steps in the chain of inequalities

$$1 - \epsilon - \delta \leq \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \tag{3.34}$$

$$= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \tag{3.35}$$

$$\leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \tag{3.36}$$

$$= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} \tag{3.37}$$

$$\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}. \tag{3.38}$$

   (c) Complete the proof of the theorem.

HISTORICAL NOTES

The Asymptotic Equipartition Property (AEP) was first stated by Shannon in his original 1948 paper [238], where he proved the result for i.i.d. processes and stated the result for stationary ergodic processes. McMillan [192] and Breiman [44] proved the AEP for ergodic finite alphabet sources. Chung [57] extended the theorem to the case of countable alphabets and Moy [197], Perez [208] and Kieffer [154] proved the $\mathscr{L}_1$ convergence when $\{X_i\}$ is continuous valued and ergodic. Barron [18] and Orey [202] proved almost sure convergence for continuous valued ergodic processes; a simple sandwich argument (Algoet and Cover [8]) will be used in Section 15.7 to prove the general AEP.