Elements of Information Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1. Asymptotic Equipartition Property
   - Convergence of Random Variables
   - Asymptotic Equipartition Property Theorem
   - Consequences of the AEP: Data Compression
   - High-Probability Sets and the Typical Set

Subsection 1

Convergence of Random Variables

Law of Large Numbers and Asymptotic Equipartition

- In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP).
- 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, for i.i.d. random variables $X_1, X_2, \ldots, X_n$, if $p(X_1, X_2, \ldots, X_n)$ is the probability of observing the sequence $X_1, X_2, \ldots, X_n$, then $\frac{1}{n}\log\frac{1}{p(X_1,X_2,\ldots,X_n)}$ is close to the entropy H.
- Thus, the probability $p(X_1, X_2, \ldots, X_n)$ assigned to an observed sequence will be close to 2^{-nH}.

Typical and Nontypical Sets of Sequences

- 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;
  - The **nontypical 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 be true with high probability;
  - Will determine the average behavior of a large sample.

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 $p(x)$, the probability of a sequence $x_1, x_2, \ldots, x_n$ is $\prod_{i=1}^n p(x_i)$.

  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.

Prediction

- 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. $\sim p(x)$.
- Apparently, we are asking for the probability of an event drawn according to the same probability distribution.
- 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

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

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

Example (Cont'd)

- In the example given, where

$$p(X_1, X_2, \ldots, X_n) = p^{\sum X_i} q^{n - \sum X_i},$$

  we are simply saying the following:
  - The number of 1's in the sequence is close to np (with high probability);
  - All such sequences have (roughly) the same probability $2^{-nH(p)}$.
- For the last statement observe that

$$\begin{array}{rcl} -nH(p) & = & -n(-p\log p - q\log q) \\ & = & \log\left(p^{np} q^{nq}\right) \\ & = & \log\left(p^{np} q^{n-np}\right). \end{array}$$

  So $p^{np} q^{n-np} = 2^{-nH(p)}$.

Convergence in Probability

- We use the following idea of convergence in probability.

Definition (Convergence of Random Variables)

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

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

Subsection 2

Asymptotic Equipartition Property Theorem

Asymptotic Equipartition Property

Theorem (AEP)

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

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

- 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,

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

Typical Sets

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 property

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

Properties of $A_\epsilon^{(n)}$

Theorem

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then
$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$
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 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;
- The number of elements in the typical set is nearly 2^{nH}.

Proof of the Theorem

(1) By the definition of $A_\epsilon^{(n)}$, if $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then

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

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

(2) By the preceding theorem, 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

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

Setting $\delta = \epsilon$, we obtain the second part of the theorem.

The identification of $\delta = \epsilon$ will conveniently simplify notation later.

Proof of the Theorem (Cont'd)

(3) We write

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

The second inequality follows from the typical set property. Hence $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

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

$$1 - \epsilon < \Pr\{A_\epsilon^{(n)}\} \leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|.$$

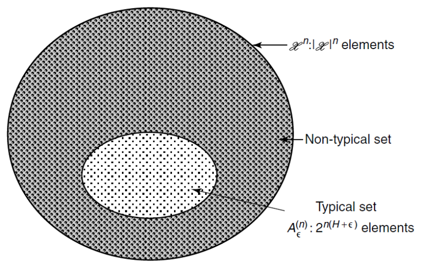
The second inequality follows again from the typical set property. Hence, $|A_\epsilon^{(n)}| \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$.

Subsection 3

Consequences of the AEP: Data Compression

Typical and Nontypical Sets of Sequences

- 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 $A_\epsilon^{(n)}$;
  - Its complement.



Encoding Scheme

- We order all elements in each set according to some order (e.g., lexicographic order).
- Then we can represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set.
  - There are $2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$.
  - So the indexing requires no more than $n(H+\epsilon)+1$ bits.
  - We prefix all these sequences by a 0.

  So total length $\leq n(H+\epsilon)+2$ bits to represent each sequence in $A_\epsilon^{(n)}$.
- Similarly, we can index each sequence not in $A_\epsilon^{(n)}$ by using not more than $n\log|\mathcal{X}|+1$ bits.
  - We prefix these indices by 1.

  We get a code for all sequences in $\mathcal{X}^n$ by using not more than $n\log|\mathcal{X}|+2$ bits.

Features of the Encoding Scheme

- The following features hold for 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 $\approx nH$.
- We use the notation x^n to denote a sequence $x_1, x_2, \ldots, x_n$.
- Let $\ell(x^n)$ be the length of the codeword corresponding to x^n.

Expected Average Length of Encoding Scheme

Theorem

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

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

for n sufficiently large.

- Thus, we can represent sequences X^n using $nH(X)$ bits on the average.
- Suppose n is sufficiently large so that $\Pr\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$.

Expected Average Length of Encoding Scheme (Cont'd)

- Then the expected length of the codeword is

$$\begin{array}{rcl} E(\ell(X^n)) & = & \sum_{x^n} p(x^n)\ell(x^n) \\ & = & \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)\ell(x^n) + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n)\ell(x^n) \\ & \leq & \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H+\epsilon)+2) \\ & & \qquad + \sum_{x^n \in A_\epsilon^{(n)c}} p(x^n)(n\log|\mathcal{X}|+2) \\ & = & \Pr\{A_\epsilon^{(n)}\}(n(H+\epsilon)+2) + \Pr\{A_\epsilon^{(n)^c}\}(n\log|\mathcal{X}|+2) \\ & \leq & n(H+\epsilon)+\epsilon(n\log|\mathcal{X}|)+2 \\ & = & n(H+\epsilon'). \end{array}$$

Note that $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small by an appropriate choice of ϵ followed by an appropriate choice of n.

Subsection 4

High-Probability Sets and the Typical Set

Smallest High-Probability Sets

- We will prove that the typical set has essentially the same number of elements as the smallest set that contains most of the probability.

Definition

For each $n = 1, 2, \ldots$, let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ be the smallest set with

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

Smallest High-Probability Sets and Typical Sets

- $B_\delta^{(n)}$ must have significant intersection with $A_\epsilon^{(n)}$.

Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, we have, for sufficiently large n,

$$\Pr\{B_\delta^{(n)}\} > 1 - \delta \quad \text{implies} \quad \frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'.$$

- 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.

Equality of First-Order in the Exponent

Definition

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

- Thus, $a_n \doteq b_n$ implies that a_n and b_n are equal to the first order in the exponent.
- We can restate the above results:

$$\text{If } \delta_n \to 0 \text{ and } \epsilon_n \to 0, \text{ then } |B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}.$$

Difference Between $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$

- Recall that a Bernoulli(θ) random variable is a binary random variable that takes on the value 1 with probability θ.
- Consider a Bernoulli sequence $X_1, X_2, \ldots, X_n$ with parameter $p = 0.9$.
- 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.
- The preceding theorem implies that:
  - $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ must both contain the sequences with about 90% 1's;
  - $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ are almost equal in size.