Asymptotic Equipartition Property

Notes on Information Theory

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

The Law of Large Numbers

- In information theory, a result 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, the sample mean is close to the expectation value, i.e.,

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

Asymptotic Equipartition Property

- The entropy is the expectation of $-\log p(X)$, since

$$H(X) = \sum p(X) \log \frac{1}{p(X)}.$$

- Let $X_{1:n}$ be independent, identically distributed (i.i.d.) random variables. For samples $x_{1:n}$ of $X_{1:n}$,

$$-\frac{1}{n} \log p(x_{1:n}) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \to E(-\log p(X))$$
$$= H(X).$$

Typical Set

- Given a distribution $p(x)$, the typical set is the set of sequences with

$$A_\epsilon^{(n)} = \{x_{1:n} | 2^{-n(H(X)+\epsilon)} \leq p(x_{1:n}) \leq 2^{-n(H(X)-\epsilon)}\}.$$

- A sequence in the typical set is a typical sequence. From above, we can see that the average log probability of a typical sequence is within ϵ of $-H(X)$.

Properties of A Typical Set

- A typical set has the following properties.

$$x_{1:n} \in A_\epsilon^{(n)} \Rightarrow -\frac{1}{n} \log p(x_{1:n}) \sim H(X) \pm \epsilon;$$

$$Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon \text{ for } n \text{ sufficiently large;}$$

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

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

- Thus the typical set has a probability of nearly 1, all elements are nearly equally likely and the total number of elements is nearly 2^{nH}.

Proof

- Property 1 follows from the definition of typical set.
- Property 2 follows from the AEP theorem.
- For property 3, note that

$$1 = \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n}) \geq \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) \geq 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}|.$$

- For property 4, note

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

Data Compression

- As a direct result of AEP, we now demonstrate that we can "compress data" to the entropy rate with a vanishing error probability.
- Specifically, we will construct a source code that
  - Use bit strings to represent each source symbol sequence.
  - The average length of the bit string per source symbol is the entropy.
  - We can reconstruct the original source symbol seqeuence from the bit string. The probability of that the reconstructed sequence is different from the original seqeunce approaches 0.

Source Code by Typical Set

- The central idea in the source code is the typical set.
  - Divide all sequences into two sets: the typical sequences and others.
  - The typical sequences can be indexed by no more than $n(H(X)+\epsilon)+1$ bits.
  - Prefix the bit string of a typical sequence by a 0-bit. This is the codeword.
  - The non-typical sequences can be indexed in $n\log|\mathcal{X}|+1$ bits. Prefix the bit string of a non-typical sequences by a 1-bit.

Code Length Per Source Symbol

- There are two important metrics for a code: the probability of error and the codeword length.
- Here we have an error-free source code as there is one-to-one correspondence between the source sequences and the codewords.
- The average number of bits for a source sequence is

$$Pr(A_\epsilon^{(n)})\left[n(H(X)+\epsilon)+2\right]+$$
$$(1-Pr(A_\epsilon^{(n)}))\left[n\log|\mathcal{X}|+2\right]\rightarrow n(H(X)+\epsilon').$$

- It follows that on average each symbol can be represented by $H(X)$ bits.

Entropy Rates

- By AEP, we are able to establish that we can describe n i.i.d. random variables in $nH(X)$ bits. But what if the random variables are dependent?
- We relax assumptions about the sources to allow them to be dependent. However, we still make the assumption of stationarity, which means the distribution is still identical.
- Under these assumptions, we will examine the average number of bits per symbol in the long run. This is called the **entropy rate**.

Stationary Stochastic Processes

- A stochastic process is an indexed sequence of random variables. It is characterized by the joint probability $p(x_1, \ldots, x_n),\ n = 1, 2, \ldots$.
- A stochastic process is **stationary** if the joint distribution is invariant with respect to a shift in the time index. That is, for all n and t,

$$Pr(X_1 = x_1, \ldots, X_n = x_n) = Pr(X_{1+t} = x_1, \ldots, X_{n+t} = x_n).$$

- The simplest kind of stationary stochastic process is the i.i.d. process.

Markov Chains

- The simplest stochastic process with dependence is one in which each random variable depends on the one preceding it and is conditionally independent of all the other preceding ones. Such a process is said to be Markov.
- A stochastic process is a Markov chain if

$$
\begin{aligned}
&Pr(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) \\
=&Pr(X_{n+1} = x_{n+1} | X_n = x_n),
\end{aligned}
$$

- The joint probability can be written as

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1) \ldots p(x_n|x_{n-1}).$$

Time-Invariant Markov Chains

- X_n is called the state at time n.
- A Markov chain is time-invariant if the state transition probability does not depend on time. Such a Markov chain can be characterized by an initial state (or distribution) and a transition probability matrix P with

$$P_{ij} = Pr(X_n = j | X_{n-1} = i),$$

which is the probability of transition from state i to state j.

Stationary Distribution

- The **stationary distribution** p of a time-invariant Markov chain is defined by

$$p(j) = \sum_i p(i) P_{ij}.$$

- If the initial state is drawn from the stationary distribution, then the Markov chain is a stationary process.
- For a "regular" Markov chain, the stationary distribution is unique and the asymptotic distribution is the stationary distribution.

A Two-State Markov Chain

- Consider a two-state Markov chain with transition probability matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

- Since there are only two states, the probability going from state 1 to state 2 must be equal to the probability going in the opposite direction in stationary situation. Thus, the stationary distribution is

$$\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$$

Entropy Rate

- If we have a stochastic process $X_1, \ldots, X_n$, a natural question to ask is how does the entropy grows with n. The entropy rate is defined as this rate of growth. Specifically,

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n),$$

  when the limit exists.
- To illustrate, we give the following examples.
  - Typewriter: $H(\mathcal{X}) = \log m$ where m is the number of equally likely output letters.
  - i.i.d. random variables: $H(\mathcal{X}) = H(X)$.

Asymptotic Conditional Entropy

- The asymptotic conditional entropy is defined by

$$H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1}),$$

  when the limit exists.
- This quantity is often easy to compute. Furthermore, it turns out that for stationary processes,

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

Proof of Existence

- We first show that $H'(\mathcal{X})$ exists for a stationary process. This follows from that $H(X_{n+1}|X_{1:n})$ is a non-increasing sequence in n

$$H(X_{n+1}|X_{1:n}) \leq H(X_{n+1}|X_{2:n}) = H(X_n|X_{1:n-1}),$$

and it is bounded from below by 0.

Proof of Equality

- To establish the equality, note

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{1:i-1})$$
$$= \lim_{n\to\infty} H(X_n | X_{1:n-1}) = H'(\mathcal{X}),$$

where we have used the theorem that

If $b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$ and $a_n \to a$, then $b_n \to a$.

Entropy Rate of Markov Chain

- The entropy rate of a stationary Markov chain is given by

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij},$$

where μ is the stationary distribution and P is the transition probability matrix.

- For example, the entropy rate of a two-state Markov chain is

$$H(\mathcal{X}) = \frac{\beta}{\alpha+\beta} H(\alpha) + \frac{\alpha}{\alpha+\beta} H(\beta).$$

Random Walk

- We will analyze a random walk on a connected weighted graph. Suppose the graph has
  - m vertices labelled by $\{1, 2, \ldots, m\}$.
  - weight $W_{ij} \geq 0$ associated with the edge from node i to node j.
  - We assume that $W_{ij} = W_{ji}$.
  - A random walk is a sequence of vertices of the graph. Given $X_n = i$, the next vertex j is chosen from the vertices connected to i with probability proportional to W_{ij}, i.e., $P_{ij} = \frac{W_{ij}}{W_i}$, where $W_i = \sum_j W_{ij}$.

Stationary Distribution

- The stationary distribution for the random walk is

$$\mu_i = \frac{W_i}{\sum_i W_i} = \frac{W_i}{2W}, \quad \text{where } W = \sum_{i,j} W_{ij}.$$

- This can be verified by checking $\mu P = \mu$, i.e.,

$$\sum_i \mu_i P_{ij} = \sum_i \frac{W_i}{2W} \frac{W_{ij}}{W_i} = \sum_i \frac{1}{2W} W_{ij} = \frac{W_j}{2W} = \mu_j$$

Entropy Rate

- The entropy rate for the random walk is

$$
\begin{aligned}
H(\mathcal{X}) = H(X_2|X_1) &= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij} \\
&= -\sum_i \frac{W_i}{2W} \sum_j \frac{W_{ij}}{W_i} \log \frac{W_{ij}}{W_i} = -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i} \\
&= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{2W} + \sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_i}{2W} \\
&= H(\ldots, \frac{W_{ij}}{2W}, \ldots) - H(\ldots, \frac{W_i}{2W}, \ldots).
\end{aligned}
$$

- If all edges are of equal weight, then

$$
H(\mathcal{X}) = \log(2E) - H(\frac{E_1}{2E}, \ldots, \frac{E_m}{2E}).
$$