1 Information Theory

1.1 Ensemble and Entropy

Definition 1. *[Ensemble] An ensemble X is a pair (A_X, P_X) where $A_X := \{x_1, ..., x_N\}$ is alphabet and $P_X := \{p_1, ..., p_N\}$ are probabilities, s.t. $\text{prob}(x_i) = p_i$ and $\sum_{i=1}^{N} p_i = 1$.*

Definition 2. *[Entropy of Ensemble] Let X an ensemble. Entropy of X is defined as*

$$H(X) := -\sum_{i=1}^{N} p_i \log_2(p_i)$$

1.2 Typical Set

Definition 3. *[Typical Set] Let X an ensemble. Given $N \in \mathbb{Z}_+$ and $\delta > 0$, then the typical set of X^N is defined as*

$$T_{N\delta} := \left\{ x \in A_{X^N} : \text{prob}\left(\left| \frac{1}{N} \log_2 \frac{1}{\text{prob}(x)} - H(X) \right| < \delta \right) \right\}.$$

Lemma 4.

1. *For $\forall x \in T_{N\delta}$, $2^{-N(H(X)+\delta)} < \text{prob}(x) < 2^{-N(H(X)-\delta)}$.*
2. *For $\forall \epsilon > 0, \delta > 0$, $\exists N > 0$, s.t. for $\forall n > N$, $\text{prob}(\{x \in T_{n\delta}\}) > 1 - \epsilon$.*

Proof. Part one is straight forward. Now prove part two in the following.

Notice that $\text{prob}(x) = \prod_{i=1}^{N} p_i$, where p_i is the probability of the i-th component of x. View $-\log_2(p_i)$ as random variable, being i.i.d. for all i, re-denoted by Y_i. Thus by center limit theorem, the probability of $\bar{Y} := (1/N)\sum_{i=1}^{N} Y_i$ obeys normal distribution, with expectation $E(Y)$ and variance $\text{Var}(Y)/\sqrt{N}$.

Recall that

$$E(Y) = \sum_s \text{prob}(y_s) y_s = \sum_{x_s \in A_X} \text{prob}(x_s)(-\log_2(p_s)) = H(X);$$

and

$$\text{Var}(Y) = \sum_s \text{prob}(y_s)(y_s - H(X))^2 = \sum_{x_s \in A_X} \text{prob}(x_s)(-\log_2(p_s) - H(X))^2$$

describing the expected derivation of $\log_2(p)$ from $H(X)$, being a finite constant, independent of N. Thus, the distribution of $(1/N)\sum_{i=1}^{N} \log_2(1/p_i)$, thus of $(1/N)\log_2(1/\text{prob}(x))$, approximates a normal distribution with expectation $H(X)$ and variance propotional to $1/\sqrt{N}$. The part one is then proved. □

Remark 5. ['Asymptotic Equipartition' Principle] We can say, without rigerousness, that almost all samples in X^N is in the typical set $T_{N\delta}$ for any given small δ as long as N is large enough. And all samples share the same probability $2^{-\text{NH}(X)}$.

1.3 The Source Coding Theorem

Theorem 6. *[Source Coding Theorem] Let X an ensemble. X^N can be compressed into more than $\text{NH}(X)$ bits with negligible risk of information loss, as $N \to \infty$; conversely if they are compressed into fewer than $\text{NH}(X)$ bits it is virtually certain that information will be lost.*

Proof. If N is large enough, then almost all message is in the typical set of X^N. There are $2^{\text{NH}(X)}$ elements in typical set, being almost equal probability. Encoding M equal probability elements needs at least $\log_2(M)$ bits, that is $\text{NH}(X)$ bits. □

1.4 The Noisy-Channel Coding Theorem