2.1. Classical Information Sources : Shannon's Noiseless Theorem

Classical information sources are emitters of letters $\boxed{x_i \in \mathcal{X}}$ from an alphabet $\mathcal{X}$ consisting of $\boxed{m = \#(\mathcal{X}) = \mathrm{card}(\mathcal{X})}$ symbols.

Any string consisting of n successive letters

$\boxed{x_{i(n)} = x_{i_1} x_{i_2} \cdots x_{i_n}}$, $x_{i_j} \in \mathcal{X}$, is a word of length n emitted by the source: $x_{i(n)} \in \mathcal{X}^{(n)} = \underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{n \text{ times}}$.
$i(n) = i_1 i_2 \cdots i_n$

There are $\#(\mathcal{X}^{(n)}) = m^n$ such words but not all of them occur with the same probability (frequency) during a speech.

Definition 2.1.1. A classical information source can be described as a stochastic process $\{X_j\}_{j=1}^{\infty}$ consisting of random variables X_j with values $x_{i_j} \in \mathcal{X}$ occurring with joint probabilities $P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = \mathrm{Prob}(X_1 = x_{i_1}, X_2 = x_{i_2}, \cdots X_n = x_{i_n})$ for all $n \geq 1$.

Remark: the probability distributions $\Pi_{[1,n]} = \{ P_{[1,n]}(x_{i(n)}) \}_{x_{i(n)} \in \mathcal{X}^{(n)}}$

must be compatible: $\sum_{i_n} P_{[1,n]}(x_{i(n)}) = \sum_{i_n} P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n})$

$$= P_{[1,n-1]}(x_{i_1} x_{i_2} \cdots x_{i_{n-1}})$$

Notice that, in general, $\sum_{i_1} P_{[1,n]}(x_{i(n)}) = \sum_{i_1} P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n})$

$$= P_{[2,n]}(x_{i_2} x_{i_3} \cdots x_{i_n})$$

$$\neq P_{[1,n-1]}(x_{i_2} x_{i_3} \cdots x_{i_n})$$

Definition 2.1.2. If a source is described by probability distributions $\Pi_{[1,n]} = \{ P_{[1,n]}(x_{i(n)}) \}_{x_{i(n)} \in \mathcal{X}^{(n)}}$

such that $\sum_{i_1} P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = P_{[1,n-1]}(x_{i_2} x_{i_3} \cdots x_{i_n})$

the source is called <u>stationary</u>.

Remark : stationary sources are such that the probability of any word depends only on the word itself and not on which use of the source it started to be emitted.

Example 2.1.1 Bernoulli sources

They are described by independent, identically distributed (i.i.d.) random variables : $$P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = \prod_{j=1}^{n} p(x_{i_j})$$

Compatibility : $$\sum_{i_n} P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = \prod_{j=1}^{n-1} p(x_{i_j}) = P_{[1,n-1]}(x_{i_1} x_{i_2} \cdots x_{i_{n-1}})$$

Stationarity : $$\sum_{i_1} P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = \prod_{j=2}^{n} p(x_{i_j}) = P_{[1,n-1]}(x_{i_2} x_{i_3} \cdots x_{i_n})$$

The letters from a Bernoulli source are all uncorrelated with each other.

Remark : given the joint probabilities $P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n})$, one defines the conditional probabilities

$$p(x_{i_n} \mid x_{i_1} x_{i_2} \cdots x_{i_{n-1}}) = \frac{P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n})}{P_{[1,n-1]}(x_{i_1} x_{i_2} \cdots x_{i_{n-1}})}$$

of the last letter on the preceding ones. Then,

$$P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = p(x_{i_n} \mid x_{i_1} \cdots x_{i_{n-1}})\, P_{[1,n-1]}(x_{i_1} x_{i_2} \cdots x_{i_{n-1}})$$

$$= p(x_{i_n} \mid x_{i_1} \cdots x_{i_{n-1}})\, p(x_{i_{n-1}} \mid x_{i_1} \cdots x_{i_{n-2}})$$

$$\times \; P_{[1,n-2]}(x_{i_1} \cdots x_{i_{n-2}})$$

$$= \left(\prod_{j=2}^{n} p(x_{i_j} \mid x_{i_1} \cdots x_{i_{j-1}}) \right) \times P_{\{1\}}(x_{i_1})$$

Example 2.1.2 Markov sources

The conditional probabilities depend only on the preceding letter:

$$p(x_{i_n} | x_{i_1} \cdots x_{i_{n-1}}) = p(x_{i_n} | x_{i_{n-1}})$$

$$P_{[1,n]}(x_{i_1} x_{i_2} \cdots x_{i_n}) = p(x_{i_n} | x_{i_{n-1}})\, p(x_{i_{n-1}} | x_{i_{n-2}}) \cdots p(x_{i_2} | x_{i_1})\, P_{\{1\}}(x_{i_1})$$

Compatibility : $\sum_{i_n} P_{[1,n]}(x_{i_1} \cdots x_{i_n}) = P_{[1,n-1]}(x_{i_1} \cdots x_{i_{n-1}}) \iff \boxed{\sum_{i_n} p(x_{i_n} | x_{i_{n-1}}) = 1}$

Stationarity : $\sum_{i_1} P_{[1,n]}(x_{i_1} \cdots x_{i_n}) = P_{[1,n-1]}(x_{i_2} \cdots x_{i_n}) \iff \boxed{\sum_{i_1} p(x_{i_2} | x_{i_1})\, P_{\{1\}}(x_{i_1})} = P_{\{1\}}(x_{i_2})$

The transition matrix $\left[T_{ij} := p(x_i | x_j)\right] =: T$ must be such that the sum of any of its columns is unity, because of compatibility.

Stationarity requires that $\boxed{T|\pi\rangle = |\pi\rangle}$ where $|\pi\rangle = \begin{pmatrix} p(x_1) \\ \vdots \\ p(x_n) \end{pmatrix}$.

Question : what is the limit to which information can be reliably compressed, that is, compressed in a manner such that it can be recovered later with arbitrarily low probability of error ?

Definition 2.1.3

A source emitting words $x_{i(n)} \in \mathcal{X}^{(n)}$ is reliably encodable at rate $R>0$ if, for any n, there exists a set $A^{(n)} \subseteq \mathcal{X}^{(n)}$ such that

$$\#(A^{(n)}) \leq 2^{nR} \quad (1)$$

and

$$\lim_n \text{Prob}(A^{(n)}) = 1 \quad (2)$$

$$\lim_n \sum_{x_{i(n)} \in A^{(n)}} P(x_{i(n)})$$

Definition 2.1.4. The information $\overset{\text{rate}}{\sqrt{H}}$ of a given source is the smallest reliable encoding rate : $H = \inf\{R : R \text{ reliable}\}$

Remark : $\boxed{\#(A^{(n)}) \leq 2^{nR}}$ means that we can specify the elements of $A^{(n)}$ by means of <u>at most R bits per symbol</u>, to be compared with the <u>$\log_2 m$ bits per symbol</u> in relation to the whole set $\mathcal{X}^{(n)}$.

$\boxed{\text{Prob}(A^{(n)}) \xrightarrow[n]{} 1}$ means that if we restrict the coding to the set $A^{(n)}$ the probability of error, that is of non coding a relevant string in $(A^{(n)})^c = \mathcal{X}^{(n)} \setminus A^{(n)}$ vanishes with $n \to +\infty$:

$$\boxed{\text{Prob}\left((A^{(n)})^c\right) = 1 - \text{Prob}(A^{(n)})} .$$

Theorem 2.1.1 For a source with $\#(\mathcal{X}) = m$ $\quad \boxed{0 \leq H \leq \log_2 m}$

Proof: $H \geq 0$ since $H = \inf\{R : R \text{ reliable}\}$ and $R > 0$.

$H \leq \log_2 m$ since $A^{(n)} \subseteq \mathcal{X}^{(n)} \Rightarrow \#(A^{(n)}) \leq 2^{n \log_2 m}$.

Exercise 2.1.1 Show how to reach $H=0$ and $H=\log_2 m$.

- $H=0$: uniform source. $p(x_{i(n)})=0$ unless

$$x_{i(n)} = x_{i(n)}^{\text{unif}} = \underbrace{x_i\, x_i \cdots x_i}_{n \text{ times}} .$$ Then, $\text{Prob}(A^{(n)} \subseteq \mathcal{X}^{(n)}) = 0$ unless

$$A^{(n)} = \left\{ x_{i(n)}^{\text{unif}} : x_i \in A \subseteq \mathcal{X} \right\} \text{ and } \boxed{\text{Prob}(A^{(n)}) = 1 \text{ iff } A = \mathcal{X}}$$

Then, $\#(A^{(n)}) = m < 2^{nR}$ for all $R>0$ with n large enough

So $\boxed{\inf R = 0}$.

- $H = \log_2 m$: equiprobable Bernoulli source.

$$p(x_{i(n)}) = \prod_{j=1}^{n} p(x_{i_j}) = \frac{1}{m^n}$$

Suppose $A^{(n)}$ is such that $\#(A^{(n)}) \le 2^{Rn}$.

Then, $$\text{Prob}(A^{(n)}) = \sum_{i^{(n)} :\, x_{i(n)} \in A^{(n)}} p(x_{i(n)}) = \frac{\#(A^{(n)})}{m^n} \le 2^{n(R-\log_2 m)} \xrightarrow[n]{} 0$$

whenever $R < \log_2 m$. Therefore, $\boxed{H = \inf_R \{R : R \text{ reliable}\} = \log_2 m}$

Lemma 2.1.1 Set $D_n(R) := \max\{\text{Prob}(A): \#(A) \le 2^{nR}\}$.

For any $\varepsilon > 0$, the information rate H satisfies

$$\lim_n D_n(H+\varepsilon) = 1 \quad (1) \qquad \text{and} \qquad H>0 \Rightarrow \lim_n D_n(H-\varepsilon) \neq 1 \quad (2)$$

Proof:

$H+\varepsilon$ is a reliable encoding rate: $\exists \{A^{(n)}\}_{n\ge 1}$ such that

$\#(A^{(n)}) \le 2^{n(H+\varepsilon)}$ and $\text{Prob}(A^{(n)}) \xrightarrow[n]{} 1$.

$1 \ge D_n(H+\varepsilon) \ge \text{Prob}(A^{(n)}) \xrightarrow[n]{} 1 \Rightarrow \lim_n D_n(H+\varepsilon) = 1$

$H>0$: $\exists\ \varepsilon>0$: $0<H-\varepsilon$ is not a reliable encoding rate

Choose $C_{H-\varepsilon}^{(n)} \subseteq \mathcal{X}^{(n)}$ such that $D_n(H-\varepsilon) = \text{Prob}\left(C_{H-\varepsilon}^{(n)}\right)$.

Then, $\#(C_{H-\varepsilon}^{(n)}) \le 2^{n(H-\varepsilon)}$ but $\lim_n \text{Prob}(C_{H-\varepsilon}^{(n)}) \neq 1$

Definition 2.1.5. The log-likelihood per source letter of $x_{i(n)}$ is

$$\xi_n(x_{i(n)}) = -\frac{1}{n}\log P(x_{i(n)}) \quad \text{if } P(x_{i(n)}) > 0$$

$$\xi_n(x_{i(n)}) = 0 \quad \text{if } P(x_{i(n)}) = 0$$

Remark: $\xi_n(x_{i(n)})$ is a random variable

Lemma 2.1.2 For all $R>0$ and $\varepsilon>0$: $\text{Prob}(\xi_n \le R) \le D_n(R) \le \text{Prob}(\xi_n \le R+\varepsilon) + 2^{-n\varepsilon}$

Proof: Set $B_R^{(n)} = \{ x_{i(n)} \in \mathcal{X}^{(n)} : P(x_{i(n)}) \ge 2^{-nR} \}$

$$= \{ x_{i(n)} \in \mathcal{X}^{(n)} : \xi_n(x_{i(n)}) = -\frac{1}{n}\log P(x_{i(n)}) \le R \}$$

- Left inequality

$$1 \ge \text{Prob}(B_R^{(n)}) = \sum_{i(n):\, x_{i(n)} \in B_R^{(n)}} P(x_{i(n)}) \ge \#(B_R^{(n)})\, 2^{-nR} \Rightarrow \#(B_R^{(n)}) \le 2^{nR}$$

Then, $D_n(R) \ge \text{Prob}(B_R^{(n)})$ since $D_n(R) = \max\{ \text{Prob}(A) : \#(A) \le 2^{nR} \}$

- Right inequality: Let $C_R^{(n)}$ be such that $\boxed{D_n(R) = \text{Prob}(C_R^{(n)})}$

$$D_n(R) = \text{Prob}(C_R^{(n)}) = \text{Prob}\left\{ x_{i^{(n)}} \in C_R^{(n)} : \xi_n(x_{i^{(n)}}) \le R+\varepsilon \right\}$$
$$+ \text{Prob}\left\{ x_{i^{(n)}} \in C_R^{(n)} : \xi_n(x_{i^{(n)}}) > R+\varepsilon \right\}$$

$$\le \text{Prob}(\xi_n \le R+\varepsilon) + \sum_{i^{(n)}:\ x_{i^{(n)}} \in C_R^{(n)},\ p(x_{i^{(n)}}) < 2^{-n(R+\varepsilon)}} p(x_{i^{(n)}})$$

$$\le \text{Prob}(\xi_n \le R+\varepsilon) + 2^{-n(R+\varepsilon)} \#(C_R^{(n)})$$

$$\le \text{Prob}(\xi_n \le R+\varepsilon) + 2^{-n(R+\varepsilon)} 2^{nR}$$

Definition 2.1.6 A sequence of random variables X_n converges in probability to a constant x ($X_n \xrightarrow{P} x$) if for any $\varepsilon > 0$ $\boxed{\lim_n \text{Prob}(|X_n - x| \ge \varepsilon) = 0}$

Example 2.1.3 Asymptotic Equipartition Property

Suppose $\xi_n \xrightarrow{P} H$, then $\forall\, \eta>0, \varepsilon>0$ there exists $n_0(\varepsilon,\eta)$ such that $n > n_0(\varepsilon,\eta)$ implies $\boxed{\text{Prob}\left(|\xi_n - H| \geq \varepsilon \right) < \eta}$

Then, $1-\eta \leq \text{Prob}\left(|\xi_n - H| < \varepsilon \right) = \text{Prob}\left\{ x_{i(n)} : H-\varepsilon < \xi_n(x_{i(n)}) < H+\varepsilon \right\}$

$$1-\eta \leq \text{Prob}\left\{ x_{i(n)} : 2^{-n(H+\varepsilon)} < p(x_{i(n)}) < 2^{-n(H-\varepsilon)} \right\}$$

Call $T^{(n)}_{\varepsilon,\eta} := \left\{ x_{i(n)} : 2^{-n(H+\varepsilon)} < p(x_{i(n)}) < 2^{-n(H-\varepsilon)} \right\}$

- $\boxed{\text{Prob}\left(T^{(n)}_{\varepsilon,\eta}\right) \geq 1-\eta}$
- $1 \geq \text{Prob}\left(T^{(n)}_{\varepsilon,\eta}\right) \geq \#\left(T^{(n)}_{\varepsilon,\eta}\right) 2^{-n(H+\varepsilon)} \Rightarrow \boxed{\#\left(T^{(n)}_{\varepsilon,\eta}\right) \leq 2^{n(H+\varepsilon)}}$
- $1-\eta \leq \text{Prob}\left(T^{(n)}_{\varepsilon,\eta}\right) \leq \#\left(T^{(n)}_{\varepsilon,\eta}\right) 2^{-n(H-\varepsilon)} \Rightarrow \boxed{\#\left(T^{(n)}_{\varepsilon,\eta}\right) \geq (1-\eta)\, 2^{n(H-\varepsilon)}}$

Remark : $T^{(n)}_{\varepsilon,\eta}$ is called a typical subset and the strings it contains typical strings. $\boxed{T^{(n)}_{\varepsilon,\eta}}$ is a high probability subset and one can decide to encode typical strings by means of roughly H bits per symbol.

Untypical strings in $\boxed{\mathcal{X}^{(n)} \setminus T^{(n)}_{\varepsilon,\eta} =: \left(T^{(n)}_{\varepsilon,\eta}\right)^c}$ can be all encoded by a same string, which would then correspond to errors. However, the probability of such an error is bounded by $\boxed{\text{Prob}\left(\left(T^{(n)}_{\varepsilon,\eta}\right)^c\right) \leq \eta}$

Theorem 2.1.2 Shannon's Noiseless Coding Theorem

If $\xi_n(x_{i(n)}) = -\frac{1}{n} \log p(x_{i(n)})$ converges to a constant ξ, then $\xi = H$, the information rate of the source.

Proof: if $\xi_n \xrightarrow{P} \xi$ then $\xi \geq 0$ since $\xi_n(x_{(n)}) \geq 0$.

From Lemma 2.1.2: $\forall\, \varepsilon > 0$

$$D_n(\xi+\varepsilon) \geq \text{Prob}(\xi_n \leq \xi+\varepsilon) \geq \text{Prob}(\xi-\varepsilon \leq \xi_n \leq \xi+\varepsilon)$$

$$= \text{Prob}(|\xi_n - \xi| \leq \varepsilon)$$

$$= 1 - \text{Prob}(|\xi_n - \xi| > \varepsilon) \xrightarrow[n]{} 1 \Longrightarrow \boxed{H \leq \xi}$$

Notice that $\boxed{\xi = 0 \Rightarrow H = 0}$.

If $\xi > 0$, for any $\varepsilon > 0$, Lemma 2.1.2 yields

$$D_n(\xi-\varepsilon) \leq \text{Prob}\left(\xi_n \leq (\xi-\varepsilon)+\frac{\varepsilon}{2}\right) + 2^{-n\varepsilon/2}$$

$$\leq \text{Prob}\left(\xi - \xi_n \geq \frac{\varepsilon}{2}\right) + 2^{-n\varepsilon/2}$$

$$\leq \text{Prob}\left(|\xi - \xi_n| \geq \frac{\varepsilon}{2}\right) + 2^{-n\varepsilon/2} \xrightarrow[n]{} 0$$

Then, Lemma 2.1.1 yields $\boxed{H \geq \xi}$