Information theory and coding

Bartek Blaszczyszyn
Notes by Antoine Groudiev

20th March 2024

Contents

Introduction

This document is Antoine Groudiev's class notes while following the class *Théorie de l'information et codage* (Information theory and coding) at the Computer Science Department of ENS Ulm. It is freely inspired by Bartek Blaszczyszyn's class notes.

This class contains two main parts: information theory, and coding theory. Information theory gives mathematical basis to build a notion of *information quantity*: given a text, how to "weight" the information contained in the sequence of characters? Some languages are more concise, but still provide the same quantity of information. Coding theory aims at finding the most concise way to represent information, with the smallest number of characters. Such theory – *source coding* – has many applications (storage, ...). Another branch of coding is *canal coding* – allowing "repetition" in a message to avoid the loss of information in a canal.

1 Entropy and source coding

We shall introduce *Shannon's entropy* of a probability distribution on a discrete space and study its basic properties. Our goal is to prove *Shannon's source coding theorem* formulated in 1948. It will allow us to interpret the entropy as a notion of the *amount of information* "carried" by random variables of a given distribution.

1.1 Shannon's entropy

Let $\mathcal{X}$ be a finite or countable set, and $p := \{p(x) \mid x \in \mathcal{X}\}$ be a probability distribution on $\mathcal{X}$.

Definition (Shannon's entropy)**.** We define (Shannon's) entropy $H(p)$ of p to be:

$$H(p) := -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{1.1.1}$$

with the convention that $0 \log 0 = 0$, and $a \log 0 = -\infty$ for $a > 0$. We will later on discuss the base of the logarithm.

Remark. *In this mathematical generalisation, $\mathcal{X}$ is the equivalent of a symbol alphabet, and p represents the text. The nature of the elements of $\mathcal{X}$ is not important: the entropy, the average, ... depend only on $\mathcal{X}$ and on the distribution p. Therefore, we can re-label the elements of $\mathcal{X}$.*

Definition (Entropy of a random variable)**.** Let X be a random variable on $\mathcal{X}$ with distribution p, that is $\mathbb{P}(X = x) = p(x)$, also denoted $X \sim p$. We define:

$$H(X) := H(p) = -\mathbb{E}(\log p(X)) \tag{1.1.2}$$

Observe that $0 \leqslant H(p) \leqslant +\infty$, and that $H(p) = 0$ if and only if X is constant almost surely.

Property 1. *Entropy is invariant with respect to deterministic injective mapping $f : \mathcal{X} \to \mathcal{Y}$:*

$$H(X) = H(f(X))$$

The entropy $H(p)$ can be interpreted as the *amount of information* carried on average by one realization from the distribution p. Later in this chapter, we shall prove a result supporting this interpretation.

Definition (Entropy units)**.** The unit of the entropy depends on the *base of the logarithm*:

- In binary basis, when $\log = \log_2$, we denote $H(p) = H_2(p)$, and its unit is the [*bit/symbol*] (per realization of X).

- In arbitrary basis $b > 0$, when $\log = \log_b$, we denote $H(p) = H_b(p)$, and its unit is the $[b - digit/symbol]$ (a b-digit is a digit which can take b values).
- In basis e, when $\log = \ln$, we denote $H(p) = H_e(p)$, and its unit is the $[nat/symbol]$ (nat is the natural unit of information).

The conversion between units can be done by changing the base of the logarithm:

$$H_b(p) = \frac{H_2(p)}{\log_2(b)}$$

Example (Bernoulli distribution). *Let $\mathcal{X} = \{0, 1\}$, and p the Bernoulli distribution such as*

$$\begin{cases} p(0) = p \\ p(1) = 1 - p \end{cases}$$

Therefore, we have $H(p) = -p \log(p) - (1 - p) \log(1 - p)$. The Bernoulli distribution with the maximum entropy is:

$$\max_{0 \leqslant p \leqslant 1} H_2(p) = H_2(1/2) = 1 \; [bit/symbol]$$

Example (Uniform distribution). *Let $\mathcal{X}$ be a finite set, and p the uniform distribution, that is:*

$$\forall x \in \mathcal{X}, \; p(x) := \frac{1}{|\mathcal{X}|}$$

Therefore, we have $H(p) = \log(|X|)$.

Example (Geometric distribution). *Let $\mathcal{X} = \mathbb{N}^\star$ and p the geometric distribution of parameter $p > 0$, that is:*

$$\forall n \in \mathbb{N}^\star, \; p(n) = p(1 - p)^{n-1}$$

Recall that $\mathbb{E}[X] = \frac{1}{p}$ when X follows a geometric law of parameter p.

Therefore, we have:

$$H(p) = \log\left(\frac{1 - p}{p}\right) - \frac{1}{p} \log(1 - p)$$

1.2 Gibbs' inequality

Theorem (Gibbs' inequality). *Let p and q be two probability distributions on $\mathcal{X}$. Then:*

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \leqslant -\sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{1.2.1}$$

Moreover, if $H(p) < \infty$, then there is equality in (1.2.1) *if and only if $p = q$.*

The right-hand-side of (1.2.1) is called *cross entropy* between p and q.

Proof. Let $x \sim p$. Gibbs' inequality is equivalent to:

$$\mathbb{E}[\log p(X)] \geqslant \mathbb{E}[\log q(X)]$$

If $\mathbb{E}[\log q(X)] = -\infty$, the inequality is trivial. Otherwise, since we have $\mathbb{E}[\log q(X)] \leqslant 0$:

$$\begin{aligned} \mathbb{E}[\log q(X)] - \mathbb{E}[\log p(X)] &= \mathbb{E}[\log q(X) - \log p(X)] \\ &= \mathbb{E}\left[\log\left(\frac{q(X)}{p(X)}\right)\right] \end{aligned}$$

log being concave, by applying Jensen's inequality, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{q(X)}{p(X)}\right)\right] &\leqslant \log \mathbb{E}\left[\frac{q(X)}{p(X)}\right] \\
&= \log \sum_{x\in\mathcal{X}} \frac{q(x)}{p(X)} p(x) \\
&= \log \sum_{x\in\mathcal{X}} q(x) \\
&= \log 1 = 0
\end{aligned}$$

The equality in Jensen's inequality holds if and only if $\frac{q(X)}{p(X)}$ is almost surely constant, that is $p = \lambda q$ almost surely; furthermore, we must have $\lambda = 1$ since both p and q are distributions, hence $p = q$ almost surely. □

Corollary (Uniform distribution maximizes entropy)**.** *Let p be a probability distribution on some set $\mathcal{X}$ with $|\mathcal{X}| < \infty$. Then:*

$$0 \leqslant H(p) \leqslant \log(|\mathcal{X}|)$$

and the equality holds if and only if p is uniform on $\mathcal{X}$.

Proof. Let $X \sim p$ and be q the uniform distribution on $\mathcal{X}$. By Gibbs' inequality:

$$H(p) \leqslant -\sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{|X|}\right) = \log|X|$$

Notice that $\log|X|$ is the entropy of the uniform distribution q. □

Corollary (Geometric distribution maximizes entropy in the set of probability measures on $\mathbb{N}^\star$ having given expectation)**.** *Let p be a probability distribution on $\mathcal{X} = \mathbb{N}^\star$ with mean $\mu = \sum_{n\geqslant 1} np(n) < \infty$. Then:*

$$H(p) \leqslant \mu\log(\mu) - (\mu-1)\log(\mu-1)$$

where the right-hand-side is the entropy of the geometric distribution with parameter $1/\mu$.

Proof. Let p be a probability distribution on $\mathcal{X} = \mathbb{N}^\star$ with mean $\mu < \infty$, and q the geometric distribution of parameter $1/\mu$. According to Gibbs' inequality,

$$\begin{aligned}
H(p) &\leqslant -\sum_{n\geqslant 1} p(n)\log q(n) \\
&= -\sum_{n\geqslant 1} p(n)\log\left(\frac{1}{\mu}\left(1-\frac{1}{\mu}\right)^{n-1}\right) \\
&= \sum_{n\geqslant 1} p(n)\log\mu - \sum_{n\geqslant 1}(n-1)p(n)\log\left(1-\frac{1}{\mu}\right) \\
&= \log\mu - \log\left(1-\frac{1}{\mu}\right)(\mu-1) \\
&= \log\mu - \left(\log(\mu-1) - \log\mu\right)(\mu-1) \\
&= \log\mu + \mu\log\mu - \mu\log(\mu-1) + \log(\mu-1) - \log\mu \\
&= \mu\log\mu - (\mu-1)\log(\mu-1) = H(q)
\end{aligned}$$

□

1.3 Entropy of random vectors

Definition (Entropy of random vectors)**.** Let $X := (X_1, \ldots, X_n)$ be a random vector on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, for some $n \geqslant 1$, with distribution

$$p(x_1^n) = p(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

The entropy of X is defined as the entropy of its distribution:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathbb{E}[\log p(X)] \tag{1.3.1}$$

Property 2 (Entropy of independent variables)**.** *Let $X := (X_1, \ldots, X_n)$ be a vector of* independent*, random variables. Then:*

$$H(X) = \sum_{i=1}^{n} H(X_i) \tag{1.3.2}$$

Proof. Let p be the joint distribution of X. By independence, $p(x) = \prod_{i=1}^{n} p_i(x_i)$. Hence:

$$\begin{aligned}
H(X) &= -\mathbb{E}[\log p(X)] \\
&= -\mathbb{E}\left[\log \prod_{i=1}^{n} p_i(X_i)\right] \\
&= -\mathbb{E}\left[\sum_{i=1}^{n} \log p_i(X_i)\right] \\
&= \sum_{i=1}^{n} -\mathbb{E}[\log p_i(X_i)] \\
&= \sum_{i=1}^{n} H(X_i)
\end{aligned}$$

□

Property 3 (Independence maximizes entropy)**.** *Let $X := (X_1, \ldots, X_n)$ be a vector of (arbitrary) random variables for some $n \geqslant 1$. Then:*

$$H(X) \leqslant \sum_{i=1}^{n} H(X_i) \tag{1.3.3}$$

Moreover, the equality holds if and only if $X_1, \ldots, X_n$ are independent.

Proof. By induction. If $n = 1$, the results holds. Let X be an n-vector of random variables and X_{n+1} another random variable. Denote $q(x, y) = p(x)p_{n+1}(y)$, where $X \sim p$ and $X_{n+1} \sim p_{n+1}$, and $(X_1, \ldots, X_n, X_{n+1}) \sim p'$. Since:

$$\begin{aligned}
H(X) + H(X_{n+1}) &= -\mathbb{E}[\log p(X) + \log p_{n+1}(X_{n+1})] \\
&= -\mathbb{E}[\log q(X, X_{n+1})] \\
&\geqslant -\mathbb{E}[\log p'(X, X_{n+1})] = H(X_1, \ldots, X_n, X_{n+1})
\end{aligned}$$

by Gibbs' inequality, the property is hereditary. Furthermore, there is equality in Gibbs' when $p' = q$, hence when X_{n+1} is independent of X, i.e. when $X_1, \ldots, X_n, X_{n+1}$ are independent. □

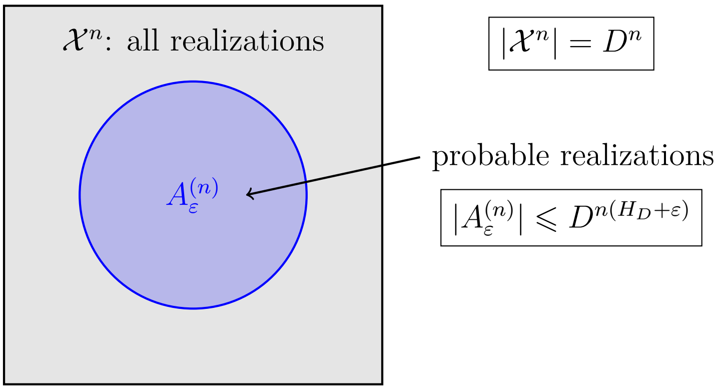
$\mathcal{X}^n$: all realizations
$|\mathcal{X}^n| = D^n$
$A_\varepsilon^{(n)}$
probable realizations
$|A_\varepsilon^{(n)}| \leqslant D^{n(H_D+\varepsilon)}$

Figure 1: Representation of $A_\varepsilon^{(n)}$

1.4 Typical sequences of random variables

Definition (Typical sequences)**.** Let $\mathcal{X}$ be a set with $D := |\mathcal{X}| < \infty$, and $X = (X_1, \dots, X_n) \in \mathcal{X}^n$ a vector of independent and identically distributed random variables. Let p be the distribution of X on $\mathcal{X}$, with $p(x) = \prod_{i=1}^n p(x_i)$. We denote $H_D := H_D(p) = -\mathbb{E}[\log_D p(X)]$, expressed in D-digits/symbol.

For $\varepsilon > 0$, the following subset of realizations of $\mathcal{X}^n$

$$A_\varepsilon^{(n)} := \left\{ x \in \mathcal{X}^n \ : \ \Big| -\frac{1}{n}\sum_{i=1}^n \log_D p(x_i) - H_D \Big| \leqslant \varepsilon \right\} \subseteq \mathcal{X}^n \tag{1.4.1}$$

is called the set of ε-typical vectors in $\mathcal{X}^n$ with respect to p.

Intuitively, the typical vectors are the vectors that probabilistically appear a lot, and that we need to represent faithfully.

Remark.

$$\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^n \log_D p(X_i) \right] = \mathbb{E}[-\log_D p(X)] = H_D$$

and, by the Law of Large Numbers (LNN for short):

$$\lim_{n\to+\infty} -\frac{1}{n}\sum_{i=1}^n \log_D p(X_i) = \mathbb{E}[-\log_D p(X)] = H_D$$

We shall see that the probability distribution of X concentrates on the set of typical sequences, and, dependending on the entropy H_D, the dimension of this set can be smaller than n (the dimension of the whole space $\mathcal{X}^n$).

Property 4 (Typical sequences concentrate probability)**.** *Let $X = (X_1, \dots, X_n)$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$, and $D := |\mathcal{X}| < \infty$. We have:*

$$\lim_{n\to+\infty} \mathbb{P}(X \in A_\varepsilon^{(n)}) = 1 \tag{1.4.2}$$

and

$$|A_\varepsilon^{(n)}| \leqslant D^{n(H_D+\varepsilon)} \tag{1.4.3}$$

Proof. (1.4.2) follows from the LLN. For (1.4.3), observe that:

$$\begin{aligned} x \in A_\varepsilon^{(n)} &\implies -\sum_{i=1}^{n} \log_D p(x_i) \leqslant n(H_D + \varepsilon) \\ &\iff \log_D \left(\prod_{i=1}^{n} p(X_i) \right) \geqslant -n(H_D + \varepsilon) \\ &\iff \log_D p(x) \geqslant -n(H_D + \varepsilon) \\ &\iff p(x) \geqslant D^{-n(H_D+\varepsilon)} \end{aligned}$$

and since:

$$\begin{aligned} 1 \geqslant \mathbb{P}(X \in A_\varepsilon^{(n)}) &= \sum_{x \in A_\varepsilon} p(x) \\ &\geqslant |A_\varepsilon^{(n)}| D^{-n(H_D+\varepsilon)} \end{aligned}$$

which completes the proof. □

Property 5 ($A_\varepsilon^{(n)}$ is the smallest set concentrating probability)**.** *Under the assumptions of Property 4, let* $B \subseteq \mathcal{X}^n$ *and* $R > 0$ *such that*

$$\lim_{n \to +\infty} \mathbb{P}(X \in B) = 1$$

and

$$|B| \leqslant D^{nR}$$

Then $R \geqslant H_D$*, that is that* $A_\varepsilon^{(n)}$ *is the smallest set concentrating probability.*

Proof. Let $\varepsilon > 0$, and assume $D > 1$, otherwise the result is trivial. Observe that:

$$\begin{aligned} x \in A_\varepsilon^{(n)} &\implies -\sum_{i=1}^{n} \log_D p(X_i) \geqslant n(H_D - \varepsilon) \\ &\iff p(x) \leqslant D^{-n(H_D-\varepsilon)} \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{P}(X \in A_\varepsilon^{(n)} \cap B) &\leqslant |B| D^{-n(H_D \varepsilon)} \\ &\leqslant D^{-n(H_D - R - \varepsilon)} \end{aligned}$$

Since $D > 1$ and $\lim_{n \to +\infty} \mathbb{P}(X \in A_\varepsilon^{(n)} \cap B) = 1$, we must have $H_D - R - \varepsilon \leqslant 0$, meaning that $H_D - \varepsilon \leqslant R$. We complete the proof by letting $\varepsilon \to 0$. □

1.5 Entropy and source coding rate – Shannon's first theorem

Definition (Encoding and decoding)**.** Let $X^n = (X_1, \dots, X_n)$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$. We call X_i *source symbols*, and the vector X^n of n source symbols is a *source message* (or *source word*, or *block of source symbols*).

Our goal is to *encode* the source message X^n consisting of n symbols using some (hopefully smaller) number of symbols in $\mathcal{X}$. That is, to represent X^n via a function $Y^m = c^{(n)}(X^n) \in \mathcal{X}^m$ for some $m \leqslant n$, in such a way that one can recover X^n from Y^m via a *decoding function* $d^{(n)}$ at least with high probability.

Definition (Compression rate)**.** The following ratio R is called *(sources) compression rate*:

$$R := \frac{m}{n} \tag{1.5.1}$$

Definition (Error probability)**.** We define the *error probability* $P_e^{(n)}$ to be:

$$P_e^{(n)} := \mathbb{P}\left(d^{(n)}\left(c^n(X^n)\right) \neq X^n\right) \tag{1.5.2}$$

which is nothing but the probability that the decoded message is different from the source message.

Theorem (Source coding theorem – Shannon's first theorem (1948))**.** *Let $X^n \in \mathcal{X}^n$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$. Denote $D := |\mathcal{X}|$ with $1 < D < \infty$. Then:*

$$\forall R > H_D(p), \quad \begin{cases} \exists c^{(n)} : \mathcal{X}^n \to \mathcal{X}^{\lceil nR \rceil} \\ \exists d^{(n)} : \mathcal{X}^{\lceil nR \rceil} \to \mathcal{X}^n \end{cases} \quad \text{such that } \lim_{n\to+\infty} P_e(n) = 0 \tag{1.5.3}$$

Furthermore,

$$\forall R < H_D(p), \quad \begin{cases} \forall c^{(n)} : \mathcal{X}^n \to \mathcal{X}^{\lceil nR \rceil} \\ \forall d^{(n)} : \mathcal{X}^{\lceil nR \rceil} \to \mathcal{X}^n \end{cases} \quad \text{we have } \lim_{n\to+\infty} P_e(n) > 0 \tag{1.5.4}$$

Proof. Let's start by proving (1.5.3). Let $R > H(p)$ and consider $0 < \varepsilon < R - H_D(p)$. For $n \geqslant 1$, let $A_\varepsilon^{(n)}$ be the set of ε-typical sequences for the distribution p, and let $f^{(n)}$ be an injection of $A_\varepsilon^{(n)}$ into $\mathcal{X}^{\lceil nR \rceil}$. Notice that such an injection exists by (1.4.3).

Let $x_\star \in X^{\lceil nR \rceil} \setminus f^{(n)}(A_\varepsilon^{(n)})$ arbitrary. Such an $x_\star$ exists since the inequality is strict in (1.4.3) (we have $H_D + \varepsilon < R$). We define the following coding function:

$$c^{(n)} := x \to \begin{cases} f^{(n)}(x) & \text{for } x \in A_\varepsilon^{(n)} \\ x_\star & \text{otherwise} \end{cases}$$

As a decoding function, consider the inverse of $f^{(n)}$ on its image $f^{(n)}(A_\varepsilon^{(n)})$, completed arbitrarily on the whole domain $\mathcal{X}^{\lceil nD \rceil}$, that is:

$$d^{(n)} := x \to \begin{cases} \left[f^{(n)}\right]^{-1}(x) & \text{if } x \in f^{(n)}(A_\varepsilon^{(n)}) \\ x_0 & \text{otherwise} \end{cases}$$

for some arbitrary $x_0 \in \mathcal{X}^n$. Finally, note that:

$$P_e^{(n)} \leqslant \mathbb{P}(X \notin A_\varepsilon^{(n)})$$

Therefore, (1.5.3) follows from (1.4.2).

The second statement, (1.5.4), follows from Property 5 considering the set:

$$B := \{x \mid d^{(n)}\left(c^n(x)\right) = x\}$$

□

Remark (Achievable compression rates)**.** *The source coding theorem – Theorem 1.5 – says that independent D-symbols emitted by a source with distribution p can be encoded asymptotically without errors using $H_D(p)$ encoding symbols per sources symbol. Note that*

$$H_D(p) \leqslant H_D(u) = \log_D(D) = 1$$

where u is the uniform distribution. The zero-error probability is approached asymptotically when increasing the length of the encoding blocks of source symbols.

Remark (Source coding rates)**.** *In general, one may use different sets of coding symbols $\mathcal{Y}$, having arbitrary number $b := |\mathcal{Y}| > 1$ of elements (set of b-digits) together with some coding and decoding functions:*

$$\begin{cases} c^{(n)} : \mathcal{X}^n \mapsto \mathcal{Y}^{n''} \\ d^{(n)} : \mathcal{Y}^{n''} \mapsto \mathcal{X}^n \end{cases}$$

In this more general scheme, the ratio:

$$R_s := \frac{\text{number of } b\text{-digits used to encode one source message}}{\text{number of source symbols in one source message}} = \frac{n''}{n} \tag{1.5.5}$$

is called (source) coding rate. *It is expressed in b-digits/(source) symbol. Considering a bijective mapping $\mathcal{X}^{n'} \mapsto \mathcal{Y}^{n''}$ with n', n'' such that $D^{n'} = b^{n''}$ in conjunction with Shannon's first theorem, it is straightforward to see that*

$$H_D(p) \log_b D = H_b(p) \text{ [}b\text{-digit/(source) symbol]}$$

is the infimum of coding rates over asymptotically error-free source coding.

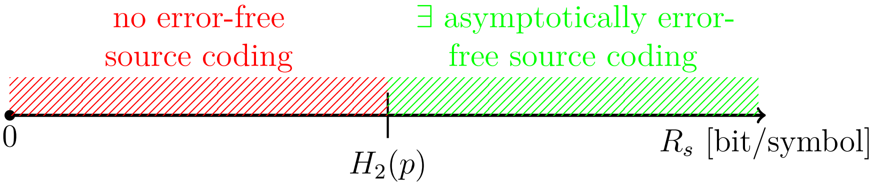
no error-free source coding
∃ asymptotically error-free source coding
0
$H_2(p)$
R_s [bit/symbol]

Figure 2: Source coding "phase transition"

2 Uniquely decipherable codes

We shall introduce some error-free source coding schemes and prove that in this class the entropy still indicates the infimum of achievable coding rates. The price to pay for error-free coding is *variable code-word length*. The main tool is Kraft's code proposed in 1949.

2.1 Uniquely decipherable and prefix-free codes

In the following, we let $\mathcal{X}$ be a finite set of *source symbols*, and $\mathcal{A}$ a set of coding symbols, called *alphabet*. We define $D := |\mathcal{A}| \in \mathbb{N}^\star$. Note that in general, $D \neq |\mathcal{X}|$. We denote $\mathcal{A}^\star$ the set of *words*, i.e. the finite (including null) sequences of character over alphabet $\mathcal{A}$. We denote $x * y$ or simply xy the concatenation between words x and y.

Definition (Code)**.** A *code* c encoding source symbols $\mathcal{X}$ in alphabet $\mathcal{A}$ is a function

$$c : \mathcal{X} \to \mathcal{A}^\star \setminus \{\varepsilon\}$$

$c(x)$ is called *code-word* of the source symbol x, and $l(x) := l(c(x))$ is the *code-word length*. The image of $\mathcal{X}$ by c, i.e. the set of all possible code-words, is called *codebook*.

In general, efficient codes will have code-words with variable length, with more probable source symbols having short code-words. In order to be able to uniquely decode a sequence of source symbols, we require more than just the injectivity of c.

Definition (Codes uniquely decipherable (UD))**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is *uniquely decipherable* (UD) if $\forall k, l \geqslant 1, \forall x_1 \dots x_k \in \mathcal{X}, \forall y_1, \dots, y_l \in \mathcal{X}$,

$$c(x_1) * \cdots * c(x_k) = c(y_1) * \cdots * c(y_l) \implies k = l \land \forall i \in [\![1, k]\!], x_i = y_i$$

Remark. *Note that c is UD if the following mapping is injective:*

$$\begin{aligned} X^\star &\to \mathcal{A}^\star \\ (x_1, \dots, x_n) &\mapsto c(x_1) * \cdots * c(x_n) \end{aligned}$$

Definition (Prefix-free (PF) codes)**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is *prefix-free (PF)* if

$$\nexists x \neq y \in \mathcal{X}, \exists a \in \mathcal{A}^\star,\ c(x) * a = c(y)$$

When such an $a \in \mathcal{A}^\star$ exists, $c(x)$ is said to be a *prefix* of $c(y)$.

Lemma 1 (PF $\implies$ UD)**.** *A prefix-free code is uniquely decipherable.*

Proof. Let c be a PF code. Suppose by contradiction that $x_1, \dots, x_k, y_1, \dots, y_l$ violate the UD condition. Let i be the smallest index such that $x_i \neq y_i$. Then, $c(x_i)$ is a prefix of $c(y_i)$ or the other way around, depending on the code length. □

Remark (Decoding v. sequential decoding of block messages)**.** *There is a difference between the decoding of UD and PF codes:*

- *A UD code allows one to uniquely decode, that is to find, given a concatenation $(a_1, \dots, a_n)$ of some code-words, the sequence of symbols x_i for $1 \leqslant i \leqslant n$ such that*

$$c(x_1) * \cdots * c(x_n) = (a_1, \dots, a_n)$$

- *A PF code allows one to decode the symbols x_i sequentially, by decoding with c the successive shortest prefixes of the encoded sequence found in the codebook. Sequential decogin simplifies decoding of blocks of symbols. As we shall see, it does not restrict the achievable performance of source coding.*

Example. *Let's consider codes $c : \mathcal{X} = \{1, 2, 3, 4\} \to \{0, 1\}^\star$. We define:*

- *A PF hence UD code of constant length, such that*

$$c(1) = (0, 0), c(2) = (0, 1), c(3) = (1, 0), c(4) = (1, 1)$$

- *A PF hence UD code of variable length*

$$c(1) = (0), c(2) = (1, 0), c(3) = (1, 10), c(4) = (1, 1, 1)$$

- *A not UD hence not PF code:*

$$c(1) = (0), c(2) = (1), c(3) = (1, 0), c(4) = (1, 1)$$

We have for instance that $c(2) * c(1) = c(3)$ and $c(2) * c(2) = c(4)$.

Property 6. *There exists codes that are UD but not PF.*

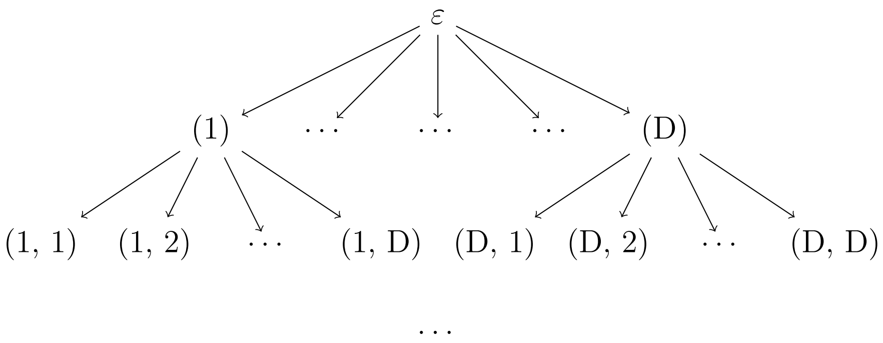

Figure 3: The entire D-ary tree

2.2 Codes on trees

Definition (k-ary tree)**.** For $k \in \mathbb{N}^\star$, a *k-ary tree* is a rooted tree in which each node (vertex) has no more than k children. A node of a k-ary tree having no children is called a *leaf*; otherwise, it is called an *intermediate node*. A tree in which each node has exactly k-children is called an *entire k-ary tree*, and is hence an infinite tree with no leaf.

Remark. *There is a natural bijection between the set of words $\mathcal{A}^\star$ expressed in the D-elements alphabet $\mathcal{A}$, and the vertices of the entire D-ary tree. Bearing in mind this bijection, we shall often identify this set of vertices with $A\star$. In particular, the empty word ε is identified with the root of the tree.*

Using the above bijection, a code $c : \mathcal{X} \to \mathcal{A}^\star$ can be seen as a mapping from $\mathcal{X}$ to the nodes of the entire D-ary tre.

Definition (Coding tree)**.** The minimal subtree of the entire D-ary tree containing the root and the code-words of the code c is called the *coding tree of c.*

Note that the coding tree is finite since $|\mathcal{X}| < \infty$, and that the leaves of the coding tree are necessarily the code-words of c, but some intermediate nodes might also be code-words.

Lemma 2 (PF coding trees)**.** *Code c is PF if and only if it is injective, and it does not have code-words at the intermediate nodes in its coding tree.*

A useful equivalent characterization of PF codes involve subtrees of the entire regular tree. For any $w \in \mathcal{A}^\star$, denote by $\mathcal{T}(w)$ the subset of elements of $\mathcal{A}^\star$ for which w is a prefix. In graph representation, $\mathcal{T}(w)$ corresponds to the subtree rooted at w and consisting of all its descendants in the entire $|\mathcal{A}|$-regular tree.

Lemma 3 (PF and subtrees)**.** *A code c is PF if and only if the subtrees of the entire regular tree rooted at the code-words corresponding to distinct source symbols are mutually disjoint:*

$$\forall x \neq y \in \mathcal{X}, \mathcal{T}(c(x)) \cap \mathcal{T}(c(y)) = \varnothing$$

Proof. $\Longrightarrow$ Assume $x \neq y \in \mathcal{X}$ and $a \in \mathcal{T}(c(x)) \cap \mathcal{T}(c(y)) \neq \varnothing$. Then both $c(x)$ and $c(y)$ are prefixes of a, and hence the shorter code-word of the two is a prefix of the longer one.

$\Longleftarrow$ If $c(x)$ is a prefix of $c(y)$ then $\mathcal{T}(c(y)) \subseteq \mathcal{T}(c(x))$, hence the intersection is not empty. □

Example (Uniform coding on the full binary tree)**.** *Assume $|\mathcal{X}| = 2^m$ for some $1 \leqslant m < \infty$, $\mathcal{A} = \{0, 1\}$. All elements of $\mathcal{X}$ can be encoded by the vertices of the m-th generation of the binary tree, or, equivalently, by the binary sequences of length m.*

Since the mapping is bijective and all code-words are equal, this is a PF and hence UD code.

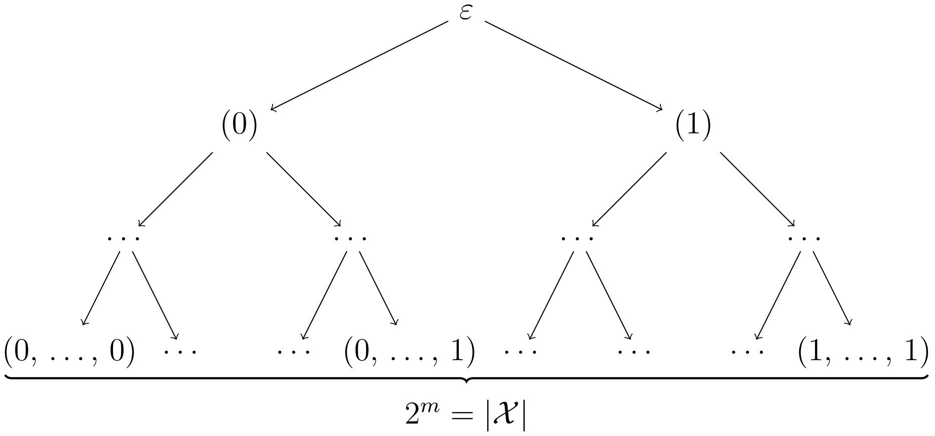

Figure 4: Binary sequences

The source coding rate is

$$R_s = \frac{\text{number of bits}}{\text{one source symbol}} = \frac{m}{1} = \log_2 |\mathcal{X}| = H_2(\text{uniform}(\mathcal{X}))$$

which corresponds to the infimum achievable coding rate when the source symbols are independently, uniformly sampled in $\mathcal{X}$. By Shannon's source coding theorem we cannot do better even accepting only asymptotically error-less coding.

A remaining question is that in case of independent source symbols sampled from some general distribution p on $\mathcal{X}$, how to achieve rates $R_S > H_2(p)$ with error-free (UD, PF) codes? The idea is to encode elemnts of $\mathcal{X}$ using binary sequences of *variable length*, with more likely source symbols having shorter code-words.

2.3 Kraft's inequality

Kraft's inequality will be our main tool in studying the performance of UD and PF codes.

Definition (Aggregated block codes)**.** For a code c and $n \in \mathbb{N}^\star$, the corresponding *aggregated block code* $c^{(n)} : \mathcal{X}^n \to \mathcal{A}^\star$ of block-length n is defined as follows:

$$\forall x \in \mathcal{X}^n, c^{(n)}(x) := c(x_1) * \cdots * c(x_n)$$

Lemma 4. *If c is UD then the aggregated block code $c^{(n)}$ is UD.*

Theorem (Kraft's inequality)**.** *Let c be a code encoding source symbols $\mathcal{X}$ with alphabet $\mathcal{A}$. Assume $D := |\mathcal{A}| < \infty$.*

1. *If c is UD then*
$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1 \tag{2.3.1}$$

2. *Let $\{l(x) \mid x \in \mathcal{X}\}$ be a collection of non-negative integers. If*
$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1$$
then there exists a PF (hence UD) code c with code-word lengths $l(x) = l(c(x))$.

Proof of Case 1. Assuming $|X| < \infty$ consider UD code c and denote by $l_{\max}$ its maximal code-word length:

$$l_{\max} = \max_{x \in \mathcal{X}} l(c(x)) < \infty$$

Let $n \in \mathbb{N}^{\star}$. Consider the aggregated block code $c^{(n)}$ corresponding to c. Recall from the preceding lemma that $c^{(n)}$ is UD. Denote by $\alpha(k)$ the number of code-words of $c^{(n)}$ of length k:

$$\alpha(k) := |\{x \in \mathcal{X} \mid l(c^{(n)}(x)) = k\}|$$

Note that $\alpha(k) = 0$ for $k = 0$ and for $k > nl_{\max}$. In general $\alpha(k) \leqslant D^k$, since $c^{(n)}$ is UD and there are at most D^k differen code-words of length k.

We have

$$\begin{aligned}\left(\sum_{x \in \mathcal{X}} D^{-l(c(x))}\right)^n &= \sum_{x \in \mathcal{X}} D^{-l(c(x_1)+\cdots+c(x_n))} \\ &= \sum_{x \in \mathcal{X}} D^{-l(c^{(n)}(x))} \\ &= \sum_{k} \alpha(k) D^{-k} \\ &\leqslant \sum_{k=1}^{nl_{\max}} D^k D^{-k} \\ &= nl_{\max}\end{aligned}$$

Consequently,

$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant (nl_{\max})^{1/n} \to_{n \to \infty} 1$$

which completes the proof of the first statement for $|\mathcal{X}| < \infty$.

If $|\mathcal{X}| = \infty$, consider an increasing sequence of subsets $(\mathcal{X}_m)$ such that $\bigcup_m \mathcal{X}_m = \mathcal{X}$. For any m, the restriction of c to $\mathcal{X}_m$ is also UD; the finite case therefore proves that

$$\sum_{x \in \mathcal{X}_m} D^{-l(c(x))} \leqslant 1$$

This being true for any m, it follows that

$$\lim_{m \to \infty} \sum_{x \in \mathcal{X}_m} D^{-l(c(x))} = \sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1$$

□

Proof of Case 2. We assume that $|\mathcal{X}| < \infty$. WLOG, let $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$, chosen such that $l(1) \leqslant \ldots \leqslant l(|\mathcal{X}|)$. Define $m := l(|\mathcal{X}|) = \max_{x \in \mathcal{X}} l(x)$. We construct a code c with code-word lengths $l(c(i)) = l(i)$ by encoding messages with some vertices of the entire D-ary tree as follows (recall $D := |\mathcal{A}|$):

- The code-word of $1 \in \mathcal{X}$ corresponds to the most recent common ancestor of the lexicographically first group of $D^{m-l(1)}$ on the m-th generation.
- By induction hypothesis, the code-word of $i \in \mathcal{X}$ corresponds to the most recent common ancestor of the lexicographically i-th group of $D^{m-l(i)}$ nodes of the m-th generation.

Note that by the assumption of the second statement,

$$\sum_{i=1}^{|\mathcal{X}|} D^{m-l(i)} \leqslant D^m$$

and thus the above construction can be completed. By construction, subtrees rooted at the selected code-words are disjoint, $\mathcal{T}(c(i) \cap \mathcal{T}(c(j))) = \varnothing$ for $1 \leqslant i \neq j \leqslant |\mathcal{X}|$, hence by lemma, the constructed code is PF.

We can similarly extend this to the case where $|\mathcal{X}| = \infty$. □

2.4 Kraft's code

We have suggested that an efficient UD code has in general variable length of the code-words. Its coding rate can be defined asymptotically on a sequence of source symbols $X_1, X_2, \dots$

$$R_s := \lim_{n\to\infty} \frac{l(c(X_1)) + \cdots + l(c(X_n))}{n} \quad \left[\frac{D-\text{digit}}{\text{source symbol}}\right] \tag{2.4.1}$$

assuming that the limit exists, thus extending the previous definition of R (1.5.1) for fixed length codes.

For $X_1, X_2, \dots$ i.i.d. with $X \sim p$, by the LLN, the limit in (2.4.1) exists a.s. and

$$R_s = \mathbb{E}[l(c(X))]$$

Therefore, looking for a rate-optimal UD code consists in minimizing the mean code-word $\mathbb{E}[l(c(X))]$. In particular, we want to know whether there exist UD codes for all compression rates $R > H_D(p)$, as it is the case for asymptotically error-free codes, according to Shannon's source coding theorem (1.5.3).

Definition (Kraft's code)**.** For a probability distribution p on $\mathcal{X}$, Kraft's code with alphabet $\mathcal{A}$ of cardinality $D := |\mathcal{A}|$ is any PF code c_K having code-words of length

$$\forall x \in \mathcal{X}, \quad l(c_K(x)) = \lceil -\log_D(p(x)) \rceil$$

Remark. *Note that*

$$\sum_{x\in\mathcal{X}} D^{-\lceil -\log_D p(x) \rceil} \leqslant \sum_{x\in\mathcal{X}} D^{\log_D p(x)} = \sum_{x\in\mathcal{X}} p(x) = 1$$

Property 7 (Mean length of Kraft's code)**.** *The mean code-word length of any Kraft's code satisfies*

$$H_D(p) \leqslant \mathbb{E}[l(c_K(X))] \leqslant H_D(p) + 1 \tag{2.4.2}$$

Furthermore, there is no UD code c with mean code-word length

$$\mathbb{E}[l(c(X))] \leqslant H_D(p)$$

Proof. For the upper bound, note that

$$\begin{aligned}\mathbb{E}[l(c_K(X))] &= \sum_{x\in\mathcal{X}} p(x) l(c(x)) \\ &\leqslant \sum_{x\in\mathcal{X}} p(x)(-\log_D p(x) + 1) \\ &= H_D(p) + 1\end{aligned}$$

The second statement follows from the second statement of Kraft's inequality combined with Gibbs' inequality. Indeed, for UD code c, by Kraft's inequality (2.3.1), $\bar{q}(x) := D^{-l(c(x))}$ can be seen as a sub-probability measure on $\mathcal{X}$. Consider a probability measure q on $\mathcal{X}$ such that $\bar{q}(x) \leqslant q(x)$. We have

$$\begin{aligned}\mathbb{E}[l(c(X))] &= \sum_{x\in\mathcal{X}} p(x) l(c(x)) \\ &= -\sum_{x\in\mathcal{X}} p(x) \log_D(D^{-l(c(x))}) \\ &\geqslant -\sum_{x\in\mathcal{X}} p(x) \log_D q(x) \\ &\geqslant -\sum_{x\in\mathcal{X}} p(x) \log_D p(x) \qquad \text{(Gibbs' inequality)} \\ &= H_D(p)\end{aligned}$$

□

Remark. *The second statement of the previous property can also be deduced from Shannon's first theorem* (1.5.4)*. Indeed, UD codes have zero-error probability and thus necessarily compression rate no smaller than $H_D(p)$. Some work is required to make this argument rigorous.*

Property 8. *The not necessarily integer-valued function*

$$l^\star(x) := -\log_D p(x)$$

minimizes $\sum_{x\in\mathcal{X}} p(x)l(x)$ within the class of functions satisfying Kraft's inequality constraint

$$\sum_{x\in\mathcal{X}} D^{-l(x)} \leqslant 1$$

Proof. We use Gibbs' inequality, noticing that $\sum_{x\in\mathcal{X}} p(x)l^\star(x) = H_D(p)$. □

Corollary (Kraft's code par block achieve all coding rates above the entropy)**.** *Let $X_1, X_2, \dots$ be i.i.d. source symbols on $\mathcal{X}$ with $X_i \sim p$. For given $n, k \in \mathbb{N}^\star$, denote by*

$$X_k^{(n)} := (X_{(k-1)n+1}, \dots, X_{kn})$$

the source messages consisting of blocks of n source symbols. Let $c_K^{(n)} : \mathcal{X}^n \to \mathcal{A}^\star$ be a Kraft's code related to the n-th product probability measure p^n, where

$$p^n(x) := p(x_1)\dots p(x_n)$$

(Note that in general, it is not the aggregated block-code related to a "one-dimensional" Kraft's code $c_k : \mathcal{X} \to \mathcal{A}^\star$, the former being used in the proof of Kraft's inequality.) The coding rate of this block-code $c^{(n)}$ is

$$\begin{aligned}
R_s &= \lim_{k\to\infty} \frac{l(c_K^{(n)}(X_1^{(n)})) + \dots + l(c_K^{(n)}(X_k^{(n)}))}{kn} \quad \left[\frac{D-\text{digit}}{\text{source symbol}}\right] \\
&= \frac{\mathbb{E}[l(c_K^{(n)}(X_1^{(n)}))]}{n} \qquad \text{(a.s. by the LLN)} \\
&\leqslant \frac{H_D(p^n)+1}{n} \qquad \text{(by (2.4.2))} \\
&= \frac{nH_D(p)+1}{n} \qquad \text{(the variables are independent)} \\
&= H_D(p) + \frac{1}{n} \to_{n\to\infty} H_D(p)
\end{aligned}$$

Remark (Entropy is the infimum of coding rates achievable over UD codes)**.** *The previous corollary shows that for any $\varepsilon > 0$ there exists a PF code (for example, Kraft's block-code) offering coding rate $R_s \leqslant H_D(p) + \varepsilon \left[\frac{D-\text{digit}}{\text{source symbol}}\right]$. Combining this with the second statement of* (2.4.2) *we conclude that $H_D(p)$ is the infimum of coding rates achievable over UD codes.*

3 Optimal codes

We shall present *Huffman's code*, introduced in 1952, which is optimal in the sens of minimizing the mean code-word length for a given source distribution.

3.1 Optimality

Let $\mathcal{X}$ be a finite set of source symbols, p a probability distribution on $\mathcal{X}$ and $\mathcal{A}$ a finite alphabet with $D := |\mathcal{A}| < \infty$ characters.

Definition (Optimal codes)**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is called an *optimal D-ary code* for p if it is a UD code which minimizes the mean code-word length $\mathbb{E}[l(c(X))]$ with $X \sim p$, in the class of UD codes:

$$\mathbb{E}[l(c(X))] = \min_{\hat{c} \text{ UD code}} \mathbb{E}[l(\hat{c}(X))]$$

Remark. *For any optimal D-ary code c for p, by 2.4.2, we have:*

$$H_D(p) \leqslant \mathbb{E}[l(c(X))] \leqslant \mathbb{E}[l(c_k(X))] \leqslant H_D(p) + 1$$

where c_k is a Kraft code (a PF code having code-words of lengths $l(x) = \lceil 1/\log p(x) \rceil$).

Property 9. *There exists an optimal code, which moreover can be taken PF.*

Proof. Assume that $p_{\min} := \min_{x\in\mathcal{X}} p(x) > 0$. The elements of $\mathcal{X}$ with zero probability can have arbitrary code-words as they do not impact the mean code length. Observe first than an optimal code exists as a solution of an optimization problem in the finite domain of the non-negative integer-valued functions $l(\cdot)$ on $\mathcal{X}$ bounded by $(H_D(p) + 1)/p_{\min} < \infty$. Indeed, $\mathbb{E}[l(c(X))] \geqslant \max_{x\in\mathcal{X}} l(c(x)) p_{\min}$, and by (3.1), no code with the mean code-word length larger than $H_D(p) + 1$ is optimal.

By Kraft's inequality (2.3.1), the code-word lengths of a given optimal code (which is UD) satisfy

$$\sum_{x\in\mathcal{X}} D^{-l(c(x))} \leqslant 1$$

By the second statement of Kraft's inequality, there exists a PF code having the same code-word lengths. It is hence optimal. □

Remark. *Combining the two statements of Kraft's inequality, a UD code c is an optimal D-ary code for p if and only if its code-word lengths $l(x) := l(c(x))$ solve the following optimization problem:*

$$\begin{cases} \text{minimize } \sum_{x\in\mathcal{X}} p(x) l(x) \\ \text{subject to } \sum_{x\in\mathcal{X}} D^{-l(x)} \leqslant 1 \end{cases} \tag{3.1.1}$$

in the set of non-negative integer valued functions l. Recall that:

- *The function $l^\star(x) := -\log p(x)$ solves* (3.1.1) *in the set of non-negative, real-valued functions yielding $\sum_{x\in\mathcal{X}} p(x) l^\star(x) = H_D(p)$.*
- *Kraft's codes assume $l(x) = \lceil -\log p(x) \rceil$, which is not necessarily a solution to* (3.1.1) *in integer-valued functions.*

3.2 Huffman's code

Let p be a probability distribution on $\mathcal{X}$, with $|\mathcal{X}| < \infty$, and consider the binary alphabet $\mathcal{A} = \{0, 1\}$.

Definition (Huffman's coding tree)**.** We construct a binary tree $\mathcal{T}_H = T_H(p)$ with vertices marked by some probabilities (real values in $[0, 1]$) executing the following algorithm:

1. Assign distinct vertices (of a binary tree to be constructed) to all elements $x \in \mathcal{X}$, *activate* and mark them by the corresponding probability $p(x)$. At this stage, all the vertices are isolated, there are no edges in the graph.

2. Take two different active vertices minimizing the sum of their probabilities. Deactivate these vertices and create a new one, being the direct common ancestor of them. Activate and mark this new vertex by the sum of the probabilities of the two vertices deactivated in this step.
3. Repeat step 2 until there is only one active vertex. Declare this vertex to be the root $\varnothing$ of the constructed graph, which is a tree. The root is marked with probability 1.

Definition (Huffman's code)**.** Huffman's code is a function $c_H : \mathcal{X} \to \{0,1\}^\star$ which assignes to $x \in \mathcal{X}$ the binary representation of the vertex corresponding to x in Huffman's binary tree $\mathcal{T}_H$.

Example (Huffman's tree and code construction)**.** *Let* $\mathcal{X} = \{1, \dots, 9\}$ *and*

$$p = (0.01, 0.02, 0.03, 0.1, 0.12, 0.2, 0.2, 0.3)$$

We obtain the following Huffman's coding tree:

Note that for the Huffman's code $c = c_H$*, the mean code length is*

$$\sum_{i=1}^{9} p(i) l(c(i)) = 2.64$$

and the entropy is

$$-\sum_{i=1}^{9} p(i) \log p(i) \simeq 2.593$$

3.3 Optimality of Huffman's code

Lemma 5 (Optimal binary code – Base condition)**.** *Let* $\mathcal{X} = \{1, \dots, n\}$*, with* $3 \leqslant n < \infty$ *and* p *be a probability distribution on* $\mathcal{X}$ *such that* $p(i) > 0$*. There exists binary PF codes*

$$c : \{1, \dots, n\} \to \{0,1\}^\star$$

optimal for p*, satisfying*

$$\begin{cases} c(n) = \omega * 0 \\ c(n-1) = \omega * 1 \text{ for } \max\limits_{i=1,\dots,n-2} p(i) \geqslant p(n-1) \geqslant p(n) > 0 \end{cases} \tag{3.3.1}$$

for some $\omega \in \{0,1\}^\star$.

Proof. Let $c : \mathcal{X} \to \{0,1\}^\star$ be an optimal binary code for p, which is PF. It exists by Property 9. Enumerate $p(1) \geqslant \dots p(n) > 0$. Then the code-word lengths of this optimal code necessarily satisfy $l(1) \leqslant \dots \leqslant l(n)$ (*The proof is left as an exercise.*)

Note that we have $l(n-1) = l(n)$. Indeed, if $l(n-1) < l(n)$, then replacing $c(n)$ by its prefix $c'(n)$ of length $l(n-1)$, we obtain a PF code that has a strictly smaller mean code-word length, thus contradicting the optimality of c.

One can also assume that $c(n) = \omega * (1)$ and $c(n-1) = \omega * (0)$. This might require exchanging some code words of length $l(n)$, preserving the PF property and the mean code-word length. □

Lemma 6 (Optimal binary code – Recursive step)**.** *Under the assumptions of Lemma 5 let a code* c *be optimal for* p *satisfying the base condition* (3.3.1)*. Then, any optimal code* c' *for the distribution* p' *defined by:*

$$\begin{cases} p'(i) = p(i), \quad \forall 1 \leqslant i \leqslant n-1 \\ p'(n-1) = p(n-1) + p(n) \end{cases}$$

makes the following code $\tilde{c}'$ *optimal for* p*:*

$$\tilde{c}'(i) := \begin{cases} c'(i) & \textit{for } 1 \leqslant i \leqslant n-2 \\ c'(n-1) * (0) & \textit{for } i = n-1 \\ c'(n) * (1) & \textit{for } i = n \end{cases}$$

Proof. Denote by L' and $\hat{L}'$ the mean code-word length of c' and $\hat{c}'$ respectively. We have

$$L' + p(n-1) + p(n) = \hat{L}'$$

Indeed,

$$\begin{aligned} L' &= \sum_{i=1}^{n-2} p(i)l(c'(i)) + (p(n-1) + p(n))l(c'(n-1)) \\ \hat{L}' &= \sum_{i=1}^{n-2} p(i)l(c'(i)) + (p(n-1) + p(n))(l(c'(n-1)) + 1) \end{aligned}$$

Consider also a code c'' shortening to $\{1, \dots, n-1\}$ the original optimal code c for p satisfying the initialization property (3.3.1):

$$c''(i) := \begin{cases} c(i) & \text{for } i = 1, \dots, n-2 \\ \omega & \text{for } i = n-1 \end{cases} \tag{3.3.2}$$

Denote by L'' the mean code of c''. We have a similar relation:

$$L'' + p(n-1) + p(n) = L$$

Therefore,

$$\begin{aligned} L &= L'' + p(n-1) + p(n) \\ &\geqslant L' + p(n-1) + p(n) \\ &= \hat{L}' \\ &\geqslant L \end{aligned}$$

thus proving that $L = \hat{L}'$ and hence that $\hat{c}'$ is optimal for p, which completes the proof. □

Corollary. *Huffman's code is optimal.*

Proof. Indeed, Huffman's code is constructed by recursion:

- It ensures the base condition (3.3.1) (when the second step of the algorithm of Huffman's binary tree is performed).
- Making the recursive step of Lemma 6 reduced the cardinality of the set of messages iteratively down to $|\mathcal{X}| = 2$. At the final step any of the two existing bijections of $\mathcal{X}$ to $\{0, 1\}$ is an optimal PF code.

Lemma 6 guarantees the optimality by the induction. □

Lemma 7. *Huffman's code minimizes the mean code-word length amongst* binary *UD codes.*

Remark. *Huffman's code minimizes the mean code-word length amongst* binary *UD codes. Comparing it to Kraft's code c_K introduced previously, using* (2.4.2) *with binary alphabet $D = 2$ we have, for $X \sim p$:*

$$H_2(p) \leqslant \mathbb{E}[l(c_H(X))] \leqslant \mathbb{E}[l(c_K(X))] \leqslant H_2(p) + 1 \tag{3.3.3}$$

Consequently, when applied per block, Huffman's code achieves asymptotically optimal coding rate converging to the entropy.

Proof of Lemma 7. Let $c : \mathcal{X} = \{1, \ldots, n\} \to \{0,1\}^\star$ be an optimal binary code for p, which is PF. It exists by Property 9. The code-word lengths of this optimal code necessarily satisfy $l(1) \leqslant \ldots \leqslant l(n)$.

Notice that we have $l(n-1) = l(n)$: indeed, if $l(n-1) < l(n)$, then replacing $c(n)$ by its prefix $c'(n)$ of length $l(n-1)$, we obtain a PF code that has a strictly small mean code-word length, thus contradicting the optimality of c.

Even if it requires exchanging some code words of length $l(n)$, preserving the PF property and the mean code-word length, one can assume that

$$\begin{cases} c(n) = \omega * (1) \\ c(n-1) = \omega * (0) \end{cases}$$

Consider c' given by (3.3.2). In what follows we shall prove that c' is optimal of p'. Denote by L and L' the mean code-word length of c and c' respectively. We have

$$L' + p(n-1) + p(n) = L$$

since

$$L = \sum_{i=1}^{n-2} p(i)l(i) + [p(n-1) + p(n)][l(\omega) + 1]$$
$$L' = \sum_{i=1}^{n-2} p(i)l(i) + [p(n-1) + p(n)]l(\omega)$$

Denote by c_0' an optimal PF code for p' and by L_0' its mean code-word length. Consider the following code $\hat{c}$ for p:

$$\hat{c}(i) = \begin{cases} c_0(i) & \text{for } i \in [\![1; n-2\ ]\!] \\ c_0(n-1) * (0) & \text{for } i = n-1 \\ c_0(n-1) * (1) & \text{for } i = n \end{cases}$$

and denote by $\hat{L}$ the mean code-word length of $\hat{c}$. That being said, we have:

$$\begin{aligned} L &= L' + p(n-1) + p(n) \\ &\geqslant L_0' + p(n-1) + p(n) \\ &= \hat{L} \\ &\geqslant L \end{aligned}$$

thus proving that $L' = L_0'$, and hence that c' is optimal for p',which completes the proof. □

3.4 Shannon-Fano-Elias code (supplement)

We shall present the Shannon-Fano-Elias code introduced in 1954. It is not optimal, but does not require ordering of the probabilities necessary to compte Huffman's code. As Kraft's and Huffman's codes, when applied per block, it achieves asymptotically optimal coding rates converging to the entropy.

4 Parsing codes

In previous lessons we have shown that coding rates close to the entropy of the source can be achieved by grouping source symbols in fixed-length blocks (messages) and coding these messages using some codes (e.g. Kraft's, Huffman's, or SFE) optimized for the block distribution. In this lesson, we shall see that equally small coding rates can be achieved by some optimal (for the given source distribution) grouping of the source symbols in *variable-length blocks*, called *parsing*, and encoding these blocks with some simple, non-optimized codes. This is the idea of Tunstall's code proposed in 1967.

4.1 Source parsing

Given a finite alphabet $\mathcal{A}$ of cardinality $D := |\mathcal{A}| < \infty$, we denote by $\mathcal{A}^\star$ the set of all finite words in alphabet $\mathcal{A}$. In what follows, we think of $\mathcal{A}$ as source symbols (previously denoted $\mathcal{X}$) and the elements of $\mathcal{A}^\star$ are possible variables-length blocks (messages) of symbols to be former by *parsing*, i.e. grouping of source symbols.

Example (Sequential source parsing using a dictionary)**.** *Parsing of the source sequence*

$$abbabbbaaababba\ldots$$

with dictionary $\mathcal{C} = \{a, ba, bba, bbb\}$ *consists in cutting the sequence successively into segments found in the dictionary:*

$$a|bba|bbb|a|a|a|ba|bba|\ldots$$

The identified segments of the source sequence will be considered as variable-length blocks (messages), to be encoded by some source code.

It is customary to interpret the process of parsing as some decoding procedure, in which initially unknow messages are recognized in the given source sequence (becoming in this "reversed" context encoded sequences). This is formalized in what follows.

Definition (Parsing code related to dictionay $\mathcal{C}$)**.** Let $\mathcal{C} := \{C(1), \ldots, C(M)\} \subset \mathcal{A}^\star \setminus \{\varepsilon\}$ be a finite subset of words, called *dictionary*, indexed by $[\![1, M]\!]$. The function

$$C : i \in [\![1, M]\!] \mapsto C(i)$$

is called the *parsing code* related to $\mathcal{C}$. We shall often identify the dictionary $\mathcal{C}$ with the corresponding parsing code.

Definition (Valid dictionary)**.** One says that dictionary $\mathcal{C} \subset \mathcal{A}^\star$ is *valid* if the following two conditions hold:

1. Each infinite sequence of characters in $\mathcal{A}$ has a prefix in $\mathcal{C}$
2. The parsing code related to $\mathcal{C}$ is PF

The first condition allows one to parse any infinite sequence of letters from $\mathcal{A}$, looking for the shortest prefixes of the remaining (not yet parsed) part in the dictionary. Second condition guarantees that this sequential parsing is the only one possible.

Example (Valid fixed-length parsing)**.** *For $n \geqslant 1$, let $\mathcal{C} := \{(x_1, \ldots, x_n) \mid x_i \in \mathcal{A}\}$. Parsing with $\mathcal{C}$ produces fixed-length blocks (messages) considered in previous lessons.*

4.2 Parsing codes and trees

Lemma 8 (Valid paring codes and complete coding trees)**.** *Let $\mathcal{C}$ be a finite dictionary over the alphabet $\mathcal{A}$ with $D := |\mathcal{A}|$. The parsing code related to $\mathcal{C}$ is valid if and only if its coding tree is a complete D-ary tree, whose leaves correspond exactly to distinct elements of $\mathcal{C}$.*

$\Longrightarrow$. Let $\mathcal{C}$ correspond to a valid code. By Lemma 2, C being PF, it has all its coding words on distinct leaves.

Let ω be an intermediate node of the coding tree, hence not a code word. Suppose by contradiction ω is missing some child, say $\omega * x$, in the coding tree. Then, any sequence starting with $\omega * x * \ldots$ does not have any prefix in $\mathcal{C}$, which contradicts the first condition of a valid parsing code. □

$\Longleftarrow$. Assume that the coding tree corresponding to $\mathcal{C}$ is complete, with $\mathcal{C}$ being exactly the leaves. The again, by Lemma 2, $\mathcal{C}$ is PF, which statisfies the second condition.

Assume that ω is the prefix of length 1 of some infinite sequence. Note that ω is in the coding tree of $\mathcal{C}$ since the root $\varepsilon \notin \mathcal{C}$ is an intermediate node. If $\omega \notin \mathcal{C}$, then ω is also an intermediate node of the coding tree of $\mathcal{C}$, and thus it has all D possible children $\omega * z$ for $z \in \mathcal{A}$, one of which is necessarily a prefix of the given sequence of length 2. The result follows by induction since the coding tree of $\mathcal{C}$ is finite. □

Lemma 9 (Node conservation law for complete trees)**.** *Let $\mathcal{C}$ be a finite dictionary of a valid parsing code. Denote by M and α, respectively, the number of leaves and intermediate nodes in the coding tree related to $\mathcal{C}$. Then:*

$$M = 1 + \alpha(D-1)$$

Proof. Let's call a *terminating node* an intermediate node whose children are leaves. By induction, we shall keep selecting one terminating node and removing all its children, until no more intermediate nodes are present in the tree. This modifies the number of intermediate nodes and leaves the following way:

$$\begin{aligned}(\alpha, M) &\longrightarrow (\alpha-1, M-D+1)\\ &\longrightarrow (\alpha-2, M-2D+2)\\ &\ldots\\ &\longrightarrow (0, M-\alpha(D-1))\end{aligned}$$

At this last step our tree consists of just one node, the root, which is also a leaf. Hence,

$$M - \alpha(D-1) = 1$$

□

4.3 Mean formulas of valid parsing

We shall perform parsing of a random sequence $X_1, X_2, \dots \in \mathcal{A}$, i.i.d., with $X_i \sim p$, and $D := |\mathcal{A}|$. We denote by W_i the i-th word identified during the sequential parsing of $X1, X_2, \ldots$ with the dictionary $\mathcal{C}$, and by $L_i := l(W_i)$ the length of this message.

Definition (Probability function)**.** We define a probability function P on finite sequences such that for any $\omega = (x_1, \ldots, x_l) \in \mathcal{A}^\star$,

$$\begin{cases} P(\omega) = \prod_{j=1}^{l} p(x_i) \\ P(\varepsilon) = 1 \end{cases}$$

Remark. *Recall that we may interpret ω as a node of the entire D-ary tree corresponding to $\mathcal{A}^\star$, and therefore*

$$\sum_{\omega \in \mathcal{A}^\star} = +\infty$$

Particularly, P is not *a probability distribution on $\mathcal{A}^\star$.*

Lemma 10 (Distribution of the parsed messages)**.** *The sequence of messages W_i obtained from i.i.d. sequences of souce symbols having distribution p by a valid parsing code C is an i.i.d sequence, with the distribution of any $W_i \sim W$ given by*

$$\mathbb{P}(W = \omega) = \begin{cases} P(\omega) & \textit{for } \omega \in \mathcal{C} \\ 0 & \textit{otherwise} \end{cases} \tag{4.3.1}$$

Proof. Left as an exercise for the reader. It is relatively easy to show that $\sum_{\omega \in \mathcal{C}} P(\omega) = 1$, and remains to show that W_i are i.i.d. □

Lemma 11 (Mean length of parsed messages)**.** *Under the assumptions of the previous Lemma, denote by $\mathcal{T}$ a complete tree of the parsing code related to $\mathcal{C}$. Then:*

$$\mathbb{E}[L] := \mathbb{E}[l(W)] = \sum_{\substack{\omega \in \mathcal{T} \\ \text{intermediate}}} P(\omega) \tag{4.3.2}$$

Proof. We shall start by proving that for $l \geqslant 0$,

$$\mathbb{P}(L > l) = \sum_{\substack{\omega \in \mathcal{T},\, l(\omega) = l \\ \text{intermediate}}} P(\omega)$$

Indeed, note by the definition of P (4.3.1) that the probability of each node is equal to the sum of the probabilities of its children:

$$P(\omega) = \sum_{x \in \mathcal{A}} P(\omega * x)$$

We can interpret the above formula as a low of *conservation of the mass* (probability) sent in the tree from the root (of mass 1) to subsequent generations in the tree $\mathcal{T}$. In this interpetation, $P(L > l)$ is the total mass received at the code words of length strictly biggen than l. Therefore, the right-hand-side of (4.3) is the total mass sent by all intermediate nodes of length l, which proves (4.3) by the mass conservation law.

Furthermore,

$$\begin{aligned} \mathbb{E}[L] &= \sum_{l=0}^{\infty} l \mathbb{P}(L = l) \\ &= \sum_{l=0}^{\infty} \mathbb{P}(L > l) \\ &= \sum_{l=0}^{\infty} \sum_{\substack{\omega \in \mathcal{T},\, l(\omega) = l \\ \text{intermediate}}} P(\omega) \\ &= \sum_{\substack{\omega \in \mathcal{T} \\ \text{intermediate}}} P(\omega) \end{aligned}$$

□

Property 10 (Parsed message entropy)**.** *Under the assumptions of the previous Lemma, with $X \sim p$, we have*

$$H(W) = H(X)\mathbb{E}[L] \tag{4.3.3}$$

Proof. Since $\mathbb{1}_{L_1 < i}$ is a function of the vector $(X_1, \ldots, X_{i-1})$, it is independent of X_i. Consequently, $\mathbb{1}_{L_1 \geqslant i}$ and X_i are also independent random variables. By Lemma 10 and (4.3.1), we have

$$\mathbb{P}(W_1 = \underbrace{(x_1, \ldots, x_l)}_{:= \omega}) = P(w) = \prod_{i=1}^{l} p(x_i)$$

Therefore,

$$\begin{aligned} H(W) = H(W_1) &= -\mathbb{E}[\log P(W_1)] \\ &= -\mathbb{E}\left[\log\left(\prod_{i=1}^{L_1} p(X_i)\right)\right] \\ &= -\mathbb{E}\left[\sum_{i=1}^{L_1} \log p(X_i)\right] \end{aligned}$$

Furthermore,

$$\begin{aligned} H(W) = -\mathbb{E}\left[\sum_{i=1}^{L_1} \log p(X_i)\right] &= -\mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{i\leqslant L_1} \log p(X_i)\right] \\ &= -\sum_{i=1}^{\infty} \mathbb{E}[ind_{i\leqslant L_1} \log p(X_i)] \\ &= -\sum_{i=1}^{\infty} \mathbb{P}(i \leqslant L_1)\mathbb{E}[\log p(X_i)] \\ &= -\mathbb{E}[\log p(X) \sum_{i=1}^{\infty} \mathbb{P}(i \leqslant L_1)] \\ &= H(X)\mathbb{E}[L_1] = H(X)\mathbb{E}[L] \end{aligned}$$

□

4.4 Parsing and translation

Parsing Recall that parsing a source sequence $x_1 * x_2 * \dots$ with a valid dictionary $\mathcal{C} = \{C(1), \dots, C(M)\}$ brakes a given source sequence into blocks of symbols (messages) $W_i = C(j_i) \in \mathcal{C}$:

$$x_1 * x_2 * \dots = W_1 * W_2 * \dots = C(j_1) * C(j_2) * \dots$$

Translation Indices $j_i \in [\![1, M]\!]$ of messages $W_i = C(j_i)$ are then encoded with a *translation code* $c : [\![1, M]\!] \to \mathcal{Y}^b$ having code-words of *fixed* length $b \geqslant 1$, and using (possibly different) alphabet $\mathcal{Y}$. Note that the number of new encoding characters $|\mathcal{Y}|$ and the code-word length b should satisfy $M \leqslant |\mathcal{Y}|^b$ so that we are guaranteed existence of a decipherable translation (injection) $c : [\![1, M]\!] \to \mathcal{Y}^b$.

The superposition of a valid parsing code $\mathcal{C}$ with an injective translation code of fixed length

$$x_1 * x_2 * \dots \longrightarrow C(j_1) * C(j_2) * \dots \longrightarrow c(j_1) * c(j_2) * \dots$$

is an injective mapping from $\mathcal{A}^\star$ into $\mathcal{Y}^\star$ and one can sequentially decode $x_1 * x_2 * \dots$ from $c(j_1) * c(j_2) * \dots$.

Example (Parsing and translating with binary alphabet)**.** *We will parse the source sequence*

$$abbabbbaaababba\dots$$

with the dictionary

$$\mathcal{C} = \{C(1) = 1, C(2) = ba, C(2) = bba, C(4) = bbb\}$$

into variable length messages, whose indices in $[\![1, 4]\!]$ *are encoded with binary translation code of length 2,* $c : [\![1, 4]\!] \to \{0, 1\}^2$.

The whole parsing and translation process maps:

$$abbabbbaaababba \dots \longrightarrow 0010110000000110\dots$$

source sequence	a	bba	bbb	a	a	a	ba	bba	$\dots$
parsing	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
message index in $\mathcal{C}$	1	3	4	1	1	1	2	3	$\dots$
translating	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
binary encoded sequence	00	10	11	00	00	00	01	10	$\dots$

Property 11 (Coding rate for parsing and translation)**.** *The coding rate for an i.i.d. sequence $X_1, X_2, \dots$ when using a superposition of parsing and translation into code-words of length b of alphabet $\mathcal{Y}$, by Lemma 10 and by the LLN, is almost surely equal to*

$$R_s := \lim_{n\to\infty} \frac{nb}{l(W_1) + \dots + l(W_n)} = \frac{b}{\mathbb{E}[L]} \left[\frac{|\mathcal{Y}|\text{-digit}}{\text{(source) symbol}}\right] \tag{4.4.1}$$

where $L = l(W)$ is the random variable having the distribution of the parsed message length.

Using (4.3.3),

$$R_s = \frac{b}{\mathbb{E}[L]} = \frac{bH(X)}{H(W)}$$

where $X \sim X_i$ and $W \sim W_i$.

Remark. *Note that the uniform entropy bound $H(W) \leqslant \log M$ and condition $M \leqslant |\mathcal{Y}|^b$ gives in general*

$$H_{\mathcal{Y}}(W) \leqslant \log_{|\mathcal{Y}|} M \leqslant b$$

impliying the lower bound on the coding rate $R_s \geqslant H_{|\mathcal{Y}|}(X)$, which is the entropy of the source, here expressed in $|\mathcal{Y}$-digits per source symbol.

The superposition of parsing and translation achieves a coding rate close to this lower bound provided

$$H_{|\mathcal{Y}|}(W) \simeq b \tag{4.4.2}$$

which can be achieved when $|\mathcal{C}| = M \simeq |\mathcal{Y}|^b$ and the distribution of the messages W on $\mathcal{C}$ is close to the uniform one:

$$H_{\mathcal{Y}} \simeq \log_{|\mathcal{Y}|} M \simeq b$$

where the first $\simeq$ holds if and only if W follows a uniform distribution of cardinality $\mathcal{C}$, and the second holds if and only if $|\mathcal{C}| = M \simeq |\mathcal{Y}|^b$.

In what follows, we shall present Tunstall's code, allowing one to achieve the previous semi-equality asymptotically when $b \to \infty$. In general, Tunstall's code transforms sequences of symbols having given distribution p into an i.i.d sequence of messages approximately uniformly distributed.

4.5 Tunstall's code

Let $\mathcal{A}$ be an alphabet with $A < |\mathcal{A}| = D < \infty$. We shall construct a valid parsing code for sequences of symbols in $\mathcal{A}$. Let $b \geqslant 1$ be a given integer; the parsed messages will undergo a translation using a translation code having code-words of fixed length b in *binary alphabet* $\mathcal{Y} = \{0, 1\}$. We seek at achieving (4.4.2) for a given probability distribution p of source symbols $X \sim p$.

Select $\alpha > 1$ such that:

$$1 + \alpha(D-1) \leqslant 2^b < 1 + (\alpha+1)(D-1) \tag{4.5.1}$$

The constructed valid parsing code will have $M = 1 + \alpha(D-1)$ code words in its dictionary $\mathcal{C}$. In fact, we construct its corresponding D-ary coding tree, which needs to be complete, with exactly M leaves corresponding to the words of the dictionary $\mathcal{C}$.

Recall from Lemma 9 (Node conservation law for valid trees) that α corresponds to the number of intermediate nodes in the coding tree of $\mathcal{C}$ that we want to construct.

Definition (Tunstall's coding tree)**.** For a given source distribution p with $p(x) > 0$ for all $x \in \mathcal{A}$, and α and b satisfiying (4.5.1), we construct a complete D-ary tree, having α intermediate nodes, by executing the following algorithm:

1. Consider complete tree $\mathcal{T}_1$ having a root and all its D children. Note that $\mathcal{T}_1$ has $\alpha_1 = 1$ intermediate node and $M_1 = D$ leaves. Mark the nodes of $\mathcal{T}_1$ by probabilities $P(\omega)$ given by (4.3.1).
2. For $n = 1, \dots, \alpha - 1$, take a leaf of $\mathcal{T}_n$ having the largest probability $P(\omega)$ and transform it into an intermediate node by adding all its children as new leaves. This gives a new complete tree $\mathcal{T}_{n+1}$ having $\alpha_{n+1} = \alpha_n + 1$ nodes and $M_{n+1} = M_n + D - 1$ leaves.
3. The dictionary of Tunstall's parsing code corresponds to the leaves of $\mathcal{T} := \mathcal{T}_\alpha$.

Theorem (Asymptotic optimality of Tunstall's code)**.** *Under the same hypothesis as the definition, let $L = l(W)$ be the length of the generic messages W parsed on the i.i.d. source sequence $X_1, X_2, \dots, X \sim p$, using Tunstall's parsing code. We have:*

$$\lim_{b \to +\infty} \frac{b}{\mathbb{E}[L]} = H_2(p) \tag{4.5.2}$$

Corollary. *By the definition of the coding rate in this case* (4.4.1)*, Tunstall's code achieves asymptotically optimal coding rate converging to the entropy*

$$R_S = \frac{b}{\mathbb{E}[L]} \xrightarrow[b \to +\infty]{} H_2(p)$$

Furthermore, by (4.3.3)*, the entropy of the block-message satisfies*

$$\lim_{b \to +\infty} \frac{H_2(W)}{b} = 1 \tag{4.5.3}$$

meaning that it is asymptotically close to $b = \log_2 2^b = H_2(\mathrm{Unif}(2^b))$, which is the entropy of a uniform random variable on $M = 2^b$ messages.

Proof. Let $\varepsilon = \omega_1, \dots, \omega_\alpha$ be the intermediate nodes created during the construction of Tunstall's coding tree. Observe the following ordering of their probabilities

$$\begin{aligned}
1 = P(\varepsilon) = P(\omega_1) & \\
&\geqslant \max_{f \text{ leaves of } \mathcal{T}_1} P(f) = P(\omega_2) \\
&\geqslant \max_{f \text{ leaves of } \mathcal{T}_2} P(f) = P(\omega_3) \\
&\dots \\
&\geqslant \max_{f \text{ leaves of } \mathcal{T}_{\alpha-1}} P(f) = P(\omega_\alpha) \\
&\geqslant \max_{f \text{ leaves of } \mathcal{T}_\alpha} P(f)
\end{aligned}$$

where the inequalities follow from

$$\forall i \in [\![1, \alpha - 1]\!], \quad \begin{cases} \{\text{leaves of } \mathcal{T}_{i+1}\} = \{\text{leaves of } \mathcal{T}_i\} \setminus \{\omega_{i+1}\} \cup \{\, \omega_{i+1} * z \mid z \in \mathcal{A}\,\} \\ P(\omega_{i+1} * z) \leqslant P(\omega_{i+1}) \end{cases}$$

Thus, we have

$$\forall i \in [\![1, \alpha]\!] \quad P(\omega_i) \geqslant P(\omega_\alpha)$$

$$\forall \text{ leaves } f \in \mathcal{T}_\alpha = \mathcal{T}, \quad P(f) \geqslant Qp_{\min}$$

where $p_{\min} = \min_{x \in \mathcal{A}} p(x) > 0$.

$$\forall \text{ leaves } f \in \mathcal{T}_\alpha = \mathcal{T}, \quad P(f) \leqslant Q$$

In particular, by the first to equations:

$$\forall \text{ leaves } f \in \mathcal{T}_\alpha = \mathcal{T}, \quad Qp_{\min} \leqslant P(f) \leqslant Q$$

Using the left-hand-side and the fact that M is the number of leaves in $\mathcal{T}$, we have

$$\begin{aligned} MQP_{\min} &\leqslant \sum_{f \text{ leaf of } \mathcal{T}} P(f) \\ &= P(\varepsilon) \\ &= 1 \end{aligned}$$

Consequently,

$$Q \leqslant \frac{1}{Mp_{\min}}$$

Therefore, using the previous result,

$$\begin{aligned} -\log_2 P(f) &\geqslant -\log_2 q \\ &\geqslant -\log_2 \left(\frac{1}{Mp_{\min}} \right) \\ &= \log_2(Mp_{\min}) \end{aligned}$$

Hence, for a random variable W having the distribution of each message parsed by the Tunstall's code (taking as its value f in the set of leaves of $\mathcal{T}$ with probability $P(f)$):

$$\begin{aligned} H_2(W) &= - \sum_{f \text{ leaf of } \mathcal{T}} P(f) \log_2 P(f) \\ &\geqslant \sum_{f \text{ leaf of } \mathcal{T}} P(f) \log_2(Mp_{\min}) \\ &= \log_2(Mp_{\min}) \end{aligned}$$

By the right-hand-side of (4.5.1), we have:

$$\begin{aligned} b &< \log_2(M + D - 1) \\ &= \log_2 \left(M \frac{p_{\min}}{P_{\min}} \left(1 + \frac{D-1}{M} \right) \right) \\ &= \log_2(Mp_{\min}) + \underbrace{\log_2 \left(\frac{1}{p_{\min}} \right) + \log_2 \left(1 + \frac{D-1}{M} \right)}_{:=c} \\ &\leqslant H_2(W) + c \\ &= \mathbb{E}[L] H_2(p) + c \end{aligned}$$

Hence

$$\frac{b}{\mathbb{E}[L]} < H_2(p) + \frac{c}{\mathbb{E}[L]}$$

On the other hand, the uniform distribution entropy bound and the right-hand-side of (4.5.1) give:

$$\begin{aligned} H_2(W) &\leqslant \log_2 M \\ &\leqslant b \end{aligned}$$

and hence, again by (4.3.3),

$$H_2(p) \leqslant \frac{b}{\mathbb{E}[L]}$$

Combining the last two results, it is enough to prove that

$$\frac{c}{\mathbb{E}[L]} \xrightarrow[b\to+\infty]{} 0$$

In this regard, note by (4.5.1) that $M = 1 + \alpha(D-1) \xrightarrow[b\to+\infty]{} +\infty$ and hence

$$c \xrightarrow[b\to+\infty]{} -\log p_{\min} \in [0, +\infty]$$

The above observation and $H_2(W) \leqslant b$ implies that $\mathbb{E}[L] \xrightarrow[b\to+\infty]{} +\infty$ which completes the proof. □

5 Universal source coding

Universal source coding consists in coding without prior knowledge of the source distribution. We shall show that for any coding rate R there exist universal codes (not depending on the source distribution) allowing for asymptotically error-free coding of any i.i.d. source having the entropy smaller than R. This is hence an extension of Shannon's first source coding theorem.

5.1 Empirical distributions

Definition (Hamming weight)**.** For $x \in \mathcal{X}^n$, with $|\mathcal{X}| < +\infty$, we define the *Hamming weight of a in x*:

$$h_x(a) := \sum_{i=1}^{n} \mathbb{1}_{x_i=a}$$

Definition (Empirical distribution associated with a vector)**.** For $x \in \mathcal{X}^n$, we define the *empirical distribution* p_x *associated with vector* x:

$$p_x(a) := \frac{h_x(a)}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=a}$$

$p_x(a)$ is the fraction of the elements of x equal to a.

Property 12. *Let $X_1, X_2, \ldots$ i.i.d. with $X_i \sim p$ on $\mathcal{X}$. Then, almost surely:*

$$p_{X_1} \xrightarrow[n\to+\infty]{\text{CVS}} p$$

that is:

$$\forall a \in \mathcal{X}, \qquad \lim_{n\to+\infty} p_{X_1}(a) = p(a)$$

almost surely.

Remark (About the Glivenko-Cantelli theorem)**.** *This consequence of the LLN is the discrete version of the Glivenko-Cantelli theorem. Property 12 says that in principle, the unknown underlying distribution p can be estimated from a given i.i.d. sequence of the random variable. There remains the question whether one can simultaneously keep learning p and encoding the sequence with a coding rate close to $H(p)$. In this lesson, we give a partial solution to this problem of universal source coding, showing that the explicit learning of p is not necessary. It is enough to know the entropy $H(p)$. Indeed, we shall show that for any given rate r there are universal block codes allowing one to obtain, for any i.i.d. source sequences having the entropy smaller than r, the decoding error that is asymptotically negligible (when the block size increases to infinity).*

Definition. Let $n \geqslant 1$.

- Let $\mathcal{P}_n$ be the collection of all empirical distributions associated to vectors of length n:
$$\mathcal{P}_n := \{\, p_x \mid x \in \mathcal{X} \,\}$$
- For $q \in \mathcal{P}_n$, let T_n be the set of all vectors of length n, whose associated empirical distribution is q:
$$T_n(q) := \{\, x \mid p_x = q \,\}$$

Lemma 12. *We have:*

$$|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \tag{5.1.1}$$

$$|\mathcal{P}_n| \leqslant (n+1)^{|\mathcal{X}|} \tag{5.1.2}$$

For $q = \left\{ \frac{k_a}{n} \,\middle|\, a \in \mathcal{X} \right\} \in \mathcal{P}_n$, we have:

$$|T_n(q)| = \frac{n!}{\prod_{a\in\mathcal{X}} k_a!} \tag{5.1.3}$$

$$(n+1)^{-|\mathcal{X}|} \cdot 2^{nH_2(q)} \leqslant |T_n(q)| \leqslant 2^{nH_2(q)} \tag{5.1.4}$$

Proof. Is left as an exercise. Note for the last equation that for $q \in \mathcal{P}_n$:

$$n^n = \left(\sum_{a\in\mathcal{X}} k_a\right)^n = \sum_{\substack{j>0 \\ \sum_{a\in\mathcal{X}} j_a = n}} \frac{n!}{\prod_{a\in\mathcal{X}} j_a!} \prod_{a\in\mathcal{X}} k_a^{j_a}$$

with the maximal term in the right-hand-side corresponding to $j_a = k_a$ for all $a \in \mathcal{X}$. □

5.2 Kullback-Leibler divergence

Definition (Kullback-Leibler (KL) divergence)**.** The *Kullback-Leibler (KL) divergence* betwween two distributions p, q on $\mathcal{X}$ is defined by:

$$D(p,q) := \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

If $\log = \log_b$, we note $D = D_b$.

Property 13 (Relationship to the cross entropy)**.** *The following relationship with the* cross entropy *$H(p,q)$ holds:*

$$H(p,q) := -\sum_{x\in\mathcal{X}} p(x) \log q(x) = H(p) + D(p,q) \tag{5.2.1}$$

Property 14 (Positivity)**.**

$$D(p,q) \geqslant 0$$
$$D(p,q) = 0 \iff p = q$$

Proof. By Gibbs' inequality. □

Remark.
- *KL divergence is not symmetric: in general, $D(p,q) \neq D(q,p)$ and thus is not a distance (metric) on the space of probability measures.*
- *If $D(p,q) < +\infty$ then p is* absolute continuous *with respect to q, denoted by $p << q$, meaning $q(x) = 0 \implies p(x) = 0$.*

Lemma 13. *Let p be a probability distribution on $\mathcal{X}$, with $|\mathcal{X}| < +\infty$. Denote by p^n the n-th product of p, that is the probability measure:*

$$p^n(x) := p(x_1)\dots p(x_n)$$

1. *For any vector $x \in \mathcal{X}$,*
$$p^n(x) = 2^{-n[(H_2(p_x)) + D_2(p_x, p)]}$$
2. *For any empirical distribution $q \in \mathcal{P}_n$:*

$$p^n(T_n(q)) \left(:= \sum_{x \in \mathcal{T}_n(q)} p^n(x) \right) \leqslant 2^{-nD_2(q,p)}$$

Proof: first statement.

$$\begin{aligned} p^n(x) &= \prod_{i=1}^{n} p(x_i) \\ &= \prod_{a \in \mathcal{X}} p(a)^{h(a,x)} \\ &= \prod_{a \in \mathcal{X}} p(a)^{np_x(a)} \\ &= 2^{\sum_{a \in \mathcal{X}} np_x(a) \log_2 p(a)} \\ &= 2^{-nH_2(p_x, p)} \\ &= 2^{-n[H_2(p_x) + D_2(p_x, p)]} \end{aligned}$$

where the last line follows from (5.2.1). □

Proof: second statement.

$$\begin{aligned} p^n(T_n(q)) &= \sum_{x \in T_n(q)} p^n(x) \\ &= |T_n(q)| \cdot 2^{-n[H_2(q) + D_2(q,p)]} \\ &\leqslant 2^{nH_2(q)} \cdot 2^{-n[H_2(q) + D_2(q,p)]} \\ &= 2^{-nD_2(q,p)} \end{aligned}$$

By positivity of the divergence and (5.2.1). □

5.3 Universal source coding with errors

In what follows, we shall prove an extension of Shannon's first theorem, by showing that all coding rates above the entropy $H(p)$ can be achieved on i.i.d. source sequences $X_1, X_2, \dots$ *without prior detailed knowledge of the underlying probability distribution* p of the source $X_i \sim p$.

For simplicity, we formulate the result for the *binary codes*, i.e. assume that the coding alphabet has two different digits $|\mathcal{A}| = 2$, say $\mathcal{A} = \{0,1\}$.

As in the proof of Shannon's first theorem, we shall construct some theoretical codes for block messages consisting of $n \geqslant 1$ source symbols encoded with blocks of size $\lceil nR \rceil$ of a binary alphabet:

$$c_n : \mathcal{X}^n \to \{0,1\}^{\lceil nR \rceil} \quad \text{and} \quad d_n : \{0,1\}^{\lceil nR \rceil} \to \mathcal{X}^n$$

thus offering the coding rate $\frac{\lceil nR \rceil}{n} \xrightarrow[n\to+\infty]{} nR$ bits per source symbol.

Unlike all previously considered UD codes, and in particular unlike Shannon's first theorem, we shall not use the distribution p to construct the coding and decoding function. Such codes are called *universal (source) codes.*

Finally, as in the proof of Shannon's first theorem, the codes will be possibly lossy, but the error probability will asymptotically go to zero when the block size n goes to $+\infty$:

$$P_\varepsilon := \mathbb{P}(d_n(c_n(X)) \neq X) \xrightarrow[n\to+\infty]{} 0$$

Theorem (Universal, asymptotically error-free source coding)**.** *Ler $R \geqslant 0$. There exist* universal *functions of the form:*

$$c_n : \mathcal{X}^n \to \{0,1\}^{\lceil nR \rceil} \quad \text{and} \quad d_n : \{0,1\}^{\lceil nR \rceil} \to \mathcal{X}^n$$

such that for any source distribution p on $\mathcal{X}$ with $H_2(p) < R$ and $X_1, X_2, \dots$ i.i.d. $X \sim p$ we have:

$$\lim_{n\to+\infty} P_\varepsilon = 0$$

Remark. *The asymptotically error-free coding of an unknown i.i.d. source offered by the above result still requires the knowledge of the source entropy $H(p)$ in order to choose a code with some rate $R > H(p)$.*

Proof. Let $R > H_2(p)$. For simplicity, assume that nR is an integer (the extension to the general case is straightforward). Following the same steps as in the proof of Shannon's first theorem, (1.5), it is enough to construct subsets $A_R^{(n)} \subseteq \mathcal{X}^n$ such that:

$$|A_R^{(n)}| < 2^{nR} \quad \text{and} \quad 1 - p^n(A_R^{(n)}) \xrightarrow[n\to+\infty]{} 0$$

The strict inequality is required only to have an extra state (denoted x_* in the proof of Theorem 1.5) to encode all elements of $\mathcal{X}^n \setminus A_R^{(n)}$.

Therefore, we define $A_R^{(n)}$ to be:

$$A_R^{(n)} := \{\, x \in \mathcal{X}^n \mid H_2(p_x) \leqslant R_n \,\} \quad \text{where} \quad R_n := R - |\mathcal{X}| \frac{\log_2(n+2)}{n}$$

Note that $\lim_{n\to+\infty} R_n = R$, and thus the sets $A_R^{(n)}$ consist of the sequences which have the entropy of the associated empirical distribution asymptotically smaller than the required rate R.

Observe that:

$$\begin{aligned}
|A_R^{(n)}| &= \left| \bigcup_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} T_n(q) \right| \\
&\leqslant \sum_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} |T_n(q)| \\
&\leqslant \sum_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} 2^{nH_2(q)} \\
&\leqslant \sum_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} 2^{nR_n} \\
&\leqslant |\mathcal{P}_n| \cdot 2^{nR_n} \\
&\leqslant (n+1)^{|\mathcal{X}|} 2^{nR_n} \\
&= 2^{nR} \left(\frac{n+1}{n+2}\right)^{|\mathcal{X}|} \\
&< 2^{nR}
\end{aligned}$$

Furthermore, note that:

$$\begin{aligned}
1 - p^n(A_R^{(n)}) &= \sum_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} p^n(T_n(q)) \\
&\leqslant \sum_{\substack{q\in\mathcal{P}_n\\ H_2(q)\leqslant R_n}} 2^{-nD_2(q,p)} \\
&\leqslant (n+1)^{|\mathcal{X}|} 2^{-n \min\{\, D_2(q,p) \mid q\in\mathcal{P}_n, H_2(q)>R_n \,\}}
\end{aligned}$$

Since $\mathcal{P}_n$ is a discret set, the min above is well-defined. Recall $\lim_{n\to+\infty} R_n = R > H_2(q)$. Consider $\varepsilon > 0$ such that $H_2(p) + \varepsilon < R$ and n large enough so that $R_n > H_2(p) + \varepsilon$. Denote by $\mathcal{P}$ all probability distributions on $\mathcal{X}$. We have:

$$\begin{aligned}
\min\{\, D_2(q,p) \mid q \in \mathcal{P}_n, H_2(q) > R_n \,\} &\geqslant \min\{\, D_2(q,p) \mid q \in \mathcal{P}_n, H_2(q) > H_2(p) + \varepsilon \,\} \\
&\geqslant \ \inf\{\, D_2(q,p) \mid q \in \mathcal{P}, H_2(q) > H_2(p) + \varepsilon \,\} =: \gamma
\end{aligned}$$

We shall argue that $\gamma > 0$. Observe that the function f such that:

$$f : q \in \mathcal{P} \longmapsto D(q,p)$$

considered as a real-valued function defined on a closed subset of $\mathbb{R}^{|\mathcal{X}|}$:

$$f : \underbrace{\left\{ q \in \mathbb{R}^{|\mathcal{X}|} \;\middle|\; \forall x \in \mathcal{X}, q(x) \geqslant 0 \wedge \sum_{x\in\mathcal{X}} q(x) = 1 \right\}}_{\text{closed subset of } \mathbb{R}^{|\mathcal{X}|}} \longrightarrow \mathbb{R}$$

is continuous. Furthermore, $\{\, q \in \mathcal{P} \mid H_2(q) \geqslant H_2(p) + \varepsilon \,\}$ interpreted as a subset of $\mathbb{R}^{|\mathcal{X}|}$, is a bounded, closed set. Since $\mathbb{R}^{|\mathcal{X}|}$ is of finite dimension, the set is compact. Thus, there exists $q_* \in \{\, q \in \mathcal{P} \mid H_2(q) \geqslant H_2(p) + \varepsilon \,\}$ such that $\gamma = D(q_*, p)$. By contradiction, suppose that $\gamma = D(q_*, p) = 0$. Necessarily, we would have $q_* = p$ and $H_2(q*) = H_2(p) \geqslant H_2(p) + \varepsilon$, thus leading to a contradiction.

Consequently,

$$1 - p^n(A_R^{(n)}) \leqslant (n+1)^{|\mathcal{X}|} \cdot 2^{-n\gamma} \quad = 2^{-n\left(\gamma - |\mathcal{X}| \frac{\log_2(n+1)}{n}\right)} \xrightarrow[n\to+\infty]{} 0$$

which completes the proof. □